Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 245

Control of Generalized Error Rates in Multiple Testing

Joseph P. Romano and Michael Wolf

May 2005

Control of Generalized Error Rates in Multiple Testing

Joseph P. Romano*
Department of Statistics
Stanford University
Stanford, CA 94305
U.S.A
E-mail: *romano@stanford.edu*

Michael Wolf
Institute for Empirical Research in Economics
University of Zurich
CH-8006 Zurich
Switzerland
E-mail: *mwolf@iew.unizh.ch*

May 18, 2005

Abstract

Consider the problem of testing s hypotheses simultaneously. The usual approach to dealing with the multiplicity problem is to restrict attention to procedures that control the probability of even one false rejection, the familiar familywise error rate (FWER). In many applications, particularly if s is large, one might be willing to tolerate more than one false rejection if the number of such cases is controlled, thereby increasing the ability of the procedure to reject false null hypotheses One possibility is to replace control of the FWER by control of the probability of k or more false rejections, which is called the k-FWER. We derive both single-step and stepdown procedures that control the k-FWER in finite samples or asymptotically, depending on the situation. Lehmann and Romano (2005a) derive some exact methods for this purpose, which apply whenever p-values are available for individual tests; no assumptions are made on the joint dependence of the p-values. In contrast, we construct methods that implicitly take into account the dependence structure of the individual test statistics in order to further increase the ability to detect false null hypotheses. We also consider the false discovery proportion (FDP) defined as the number of false rejections divided by the total number of rejections (and defined to be 0 if there are no rejections). The false discovery rate proposed by Benjamini and Hochberg (1995) controls $E(\text{FDP})$. Here, the goal is to construct methods which satisfy, for a given γ and α, $P\{\text{FDP} > \gamma\} \leq \alpha$, at least asymptotically.

KEY WORDS: Bootstrap, False Discovery Proportion, False Discovery Rate, Generalized Familywise Error Rates, Multiple Testing, Stepdown Procedure.

*Research supported by National Science Foundation grant DMS 010392.

1 Introduction

The main goal of this paper is to show how computer-intensive methods can be used to construct asymptotically valid tests of multiple hypotheses under very weak conditions. In particular, we construct computationally feasible methods which provide control (at least asymptotically) of some generalized notions of the familywise error rate. However, the theory also applies to exact finite sample control in some situations.

Consider the problem of testing hypotheses $H_1, \ldots, H_s$. A classical approach to dealing with the multiplicity problem is to restrict attention to procedures that control the probability of one or more false rejections, which is called the familywise error rate (FWER). Here the term "family" refers to the collection of hypotheses $H_1, \ldots, H_s$ that is being considered for joint testing. For a given family, control of the FWER at (joint) level α requires that FWER $\leq \alpha$ for all possible distributions of the data considered in the model, and therefore for all possible constellations of true and false hypotheses. A broad treatment of methods that control the FWER is given in Hochberg and Tamhane (1987).

Of course, safeguards against false rejections are not the only concern of multiple testing procedures. Corresponding to the power of a single test one must also consider the ability of a procedure to detect departures from the null hypotheses. When the number of tests s is large, such as in genomics studies, control of the FWER at conventional levels becomes so stringent that individual departures from the null hypotheses have little chance of being detected. For this reason, we shall consider alternatives to the FWER that control false rejections less severely so that better power can be obtained.

First, we shall consider the k-FWER, the probability of rejecting at least k true null hypotheses. Such an error rate with $k > 1$ is appropriate when one is willing to tolerate a given number of false rejections. More formally, suppose data X is available from some model $P \in \Omega$. A general hypothesis H can be viewed as a subset ω of Ω. For testing $H_i : P \in \omega_i$, $i = 1, \ldots, s$, let $I(P)$ denote the set of true null hypotheses when P is the true probability distribution; that is, $i \in I(P)$ if and only if $P \in \omega_i$. Then, the k-FWER, which depends on P is defined to be

$$k\text{-FWER} = k\text{-FWER}_P = P\{\text{reject at least } k \text{ hypotheses } H_i : i \in I(P)\} \ . \tag{1}$$

Control of the k-FWER requires that k-FWER $\leq \alpha$ for all P; that is,

$$k\text{-FWER}_P \leq \alpha \quad \text{for all } P \ . \tag{2}$$

Evidently, the case $k = 1$ reduces to control of the usual FWER.

We will also consider control of the *false discovery proportion* (FDP), defined as the total number of false rejections divided by the total number of rejections (and equal to 0 if there are no rejections). Given a user specified value $\gamma \in [0, 1)$, the measure of error control we wish to control is $P\{\text{FDP} > \gamma\}$; thus, we wish to construct methods satisfying

$$P\{\text{FDP} > \gamma\} \leq \alpha \quad \text{for all } P \ . \tag{3}$$

We will derive methods where this is (at least asymptotically) bounded by α. Evidently, control of the FDP with $\gamma = 0$ reduces to the usual FWER. Control of the false discovery rate (FDR) requires that $E(\text{FDP}) \leq \alpha$.

Recently, there have been a number of methods that control generalized error rates which are less stringent than the FWER. A prominent such technique is the FDR controlling method of Benjamini and Hochberg (1995). Additional methods that control the FDR are given in Benjamini and Yekutieli (2001) and Sarkar (2002). Genovese and Wasserman (2004) study asymptotic procedures that control the FDP (and the FDR) in the framework of a random effects mixture model. These ideas are extended in Perone Pacifico et al. (2004), where in the context of random fields, the number of null hypotheses is uncountable. Korn et al. (2004) provide methods that control both the k-FWER and FDP; they provide some justification for their methods, but they are limited to a multivariate permutation model. Alternative methods of control of the k-FWER and FDP are given in van der Laan et al. (2004); they include both finite sample and asymptotic results. Like the present work, their work attempts to capture the dependence between the tests with the goal of improved ability to detect false hypotheses; comparisons between the methods will be made later; see Section 5.

Some existing methods that control the k-FWER and FDP are we now briefly reviewed. Suppose that p-values $\hat{p}_1, \ldots, \hat{p}_s$ are available for testing $H_1, \ldots, H_s$. Formally, for $\hat{p}_i$ to be a p-value, it is required that, for all $u \in [0, 1]$ and all $P \in \omega_i$,

$$P\{\hat{p}_i \leq u\} \leq u \ . \tag{4}$$

Then, for any fixed k, the procedure that rejects H_i if $\hat{p}_i \leq k\alpha/s$ controls the k-FWER at level α, and can be viewed as a generalization of the Bonferroni procedure which uses $k = 1$; see Lehmann and Romano (2005a). It is an example of a *single-step* procedure, meaning any null hypothesis is rejected if its corresponding p-value is less than or equal to a common cutoff value.

Improvements are possible by considering a class of *stepdown* procedures, which we now describe. Order the p-values by

$$\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(s)} \ ,$$

and let $H_{(1)}, \ldots, H_{(s)}$ denote the corresponding hypotheses. Let

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s \tag{5}$$

be constants. If $\hat{p}_{(1)} > \alpha_1$, reject no null hypotheses. Otherwise, if

$$\hat{p}_{(1)} \leq \alpha_1, \ldots, \hat{p}_{(r)} \leq \alpha_r \ , \tag{6}$$

reject hypotheses $H_{(1)}, \ldots, H_{(r)}$ where the largest r satisfying (6) is used. That is, a stepdown procedure starts with the most significant p-value and continues rejecting hypotheses as a remaining p-value is deemed "small", where "small" is determined by the critical value α_j at step j. The procedure of Holm (1979) uses $\alpha_j = \alpha/(s - j + 1)$ and controls the FWER at

level α. For general k, consider the following generalized Holm stepdown procedure described in (6), where now we specifically set

$$\alpha_j = \begin{cases} \frac{k\alpha}{s} & j \leq k \\ \frac{k\alpha}{s+k-j} & j > k \end{cases} \tag{7}$$

Of course, the α_j depend on s and k, but we suppress this dependence in the notation. Then, the stepdown method described in (6) with α_j given by (7) controls the k-FWER; that is, (2) holds; see Hommel and Hoffman (1987) and Lehmann and Romano (2005a).

Turning to FDP control, Lehmann and Romano (2005a) reasoned as follows. To develop a stepdown procedure satisfying (3), let F denote the number of false rejections. At step j, having rejected $j-1$ hypotheses, we want to guarantee $F/j \leq \gamma$, i.e. $F \leq \lfloor \gamma j \rfloor$, where $\lfloor x \rfloor$ is the greatest integer $\leq x$. So, if $k = \lfloor \gamma j \rfloor + 1$, then $F \geq k$ should have probability no greater than α; that is, we must control the number of false rejections to be $\leq k$. Therefore, we use the stepdown constant α_j with this choice of k (which now depends on j); that is,

$$\alpha_j = \frac{(\lfloor \gamma j \rfloor + 1)\alpha}{s + \lfloor \gamma j \rfloor + 1 - j} \ . \tag{8}$$

Under certain dependence assumptions on the p-values, this method satisfies (3). Similar methods that hold under no dependence assumptions are developed in Lehmann and Romano (2005a), Romano and Shaikh (2004) and Romano and Shaikh (2005).

In general, these generalized Holm type of methods assume a least favorable joint distribution for the p-values. In contrast, here we implicitly try to estimate the joint distribution of p-values with the hopes of greater ability to detect false hypotheses.

In Section 2, we discuss stepdown methods that control the k-FWER in finite samples. Such methods proceed stepwise by testing intersection hypotheses at each step. Using a simple monotonicity condition for critical values, it is shown how computationally feasible (but possibly computer-intensive) methods can be constructed.

For any $K \subset \{1, \ldots, s\}$, let H_K denote the hypothesis that all H_i with $i \in K$ are true. The closure method of Marcus et al. (1976) allows one to construct methods that control the FWER if one knows how to test each intersection hypothesis H_K. Indeed, this method can be generalized to control the k-FWER; see Appendix A. However, in general, this might require the construction of nearly 2^s tests. The constructions studied here only require a much lower order number of tests; for example, the number of such tests is of order s in Algorithm 2.2. In fact, the monotonicity assumptions we invoke can be viewed as justification to achieve this much lower order. (In some cases, shortcuts to applying the closure method are known. For example, Westfall et al. (2001) show how to apply closure to Fisher combination tests with only s^2 evaluations.)

In general, we suppose that rejection of H_i is based on large values of a test statistic $T_{n,i}$. (To be consistent with later notation, the n is used for asymptotic purposes and typically refers to sample size.) Of course, if a p-value $\hat{p}_i$ is available for testing H_i, one possibility is to take $T_{n,i} = -\hat{p}_i$. Then, we restrict attention to tests that reject an intersection hypothesis

H_K when the kth largest of the test statistics $\{T_{n,i} : i \in K\}$ is large. In some problems where a monotonicity condition holds (distinct from the monotonicity assumption here), Lehmann et al. (2005), for the particular case of $k = 1$, show that such stepwise procedures are optimal in a maximin sense. In other situations, it may be better to consider other test statistics that combine the individual test statistics in a more powerful way. A related issue is one of balance; see Remark 3.5. At this time, our primary goal is to show how stepdown procedures can be constructed quite generally that control the k-FWER and FDP under minimal conditions; in particular, we do not have to assume the subset pivotality condition of Westfall and Young (1993, page 42).

In Section 2, we show that, if we estimate critical values that have a monotonicity property, then the basic problem of constructing a valid multiple test procedure that controls the k-FWER can essentially be reduced to the problem of sequentially constructing critical values for (at most order s) single tests that control the usual Type 1 error. In particular, if finite sample methods which offer control of the Type 1 error are available for each of the individual tests, then this will immediately translate into control of the k-FWER. For example, this allows us to directly apply what we know about tests based on permutation and randomization distributions. Alternatively, we can apply bootstrap and subsampling methods to achieve asymptotic control, as described in Section 3. Results for control of the FDP are obtained in Section 4. Comparisons with augmentation procedures are discussed in Section 5. In Section 6, we present a simulation study to examine the finite sample performance of some of the methods we suggest. All proofs are collected in an appendix.

2 Basic Results for Control of the k-FWER

Suppose data X is generated from some unknown probability distribution P. In anticipation of asymptotic results, we may write $X = X^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions Ω, though we make no rigid requirements for Ω. Indeed, Ω may be a nonparametric model, a parametric model, or a semiparametric model.

Consider the problem of simultaneously testing a hypothesis H_i against H_i', for $i = 1, \ldots, s$. Of course, a hypothesis H_i can be viewed as a subset, ω_i, of Ω, in which case the hypothesis H_i is equivalent to $P \in \omega_i$ and H_i' is equivalent to $P \notin \omega_i$. For any subset $K \subset \{1, \ldots, s\}$, define

$$H_K = \bigcap_{i \in K} H_i$$

to be the *intersection* hypothesis that $P \in \bigcap_{i \in K} \omega_i$.

Suppose that a test of the individual hypothesis H_i is based on a test statistic $T_{n,i}$, with large values indicating evidence against H_i. For an individual hypothesis, numerous approaches exist to approximate a critical value, such as those based on classical likelihood theory, bootstrap tests, Edgeworth expansions, permutation tests, etc.. The main problem addressed in the present work is to construct procedures that control generalized familywise error rates, the k-FWER and FDP, in finite samples or at least asymptotically.

Some further notation is required. Suppose $\{y_i : i \in K\}$ is a collection of real numbers indexed by a finite set K having $|K|$ elements. Then, for $k \leq |K|$, the k-max$(y_i : i \in K)$ is used to denote the kth largest value of the y_i with $i \in K$. So, if the elements y_i, $i \in K$, are ordered as

$$y_{(1)} \leq \cdots \leq y_{(|K|)} \ ,$$

then

$$k\text{-max}(y_i : i \in K) = y_{(|K|-k+1)} \ .$$

2.1 Single-step Control of the k-FWER

Throughout this section, k is fixed. First, we briefly discuss a single-step approach to control of the k-FWER, since it serves as a building block for the more powerful stepdown procedures considered later, much in the same way the Bonferroni method is a building block for the more powerful Holm method. For any subset $K \subset \{1, \ldots, s\}$, let $c_{n,K}(\alpha, k, P)$ denote an α-quantile of the distribution of k-max$(T_{n,i} : i \in K)$ under P. Concretely,

$$c_{n,K}(\alpha, k, P) = \inf\{x : \ P\{k\text{-max}(T_{n,i} : i \in K) \leq x\} \geq \alpha\} \ . \tag{9}$$

(We use the subscript n for asymptotic purposes later on, though the priority in this section is to study nonasymptotic results.)

For testing the intersection hypothesis H_K with $K \subset \{1, \ldots, s\}$, it is only required to approximate a critical value for $P \in \bigcap_{i \in K} \omega_i$. Because there may be many such P, we define

$$c_{n,K}(1-\alpha, k) = \sup\{c_{n,K}(1-\alpha, k, P) : \ P \in \bigcap_{i \in K} \omega_i\} \ . \tag{10}$$

(In order to define $c_{n,K}(\alpha, k)$, we implicitly assumed $\bigcap_{i=1}^{s} \omega_i$ is not empty.)

Consider the idealized test that rejects any H_i for which $T_{n,i} > c_{n,I(P)}(1-\alpha, k, P)$. This is a single-step method in that each $T_{n,i}$ is compared with a common cutoff. However, this is an idealization because the critical value $c_{n,I(P)}(1-\alpha, k, P)$ is in general unknown. Such a fictional test clearly controls the k-FWER at level α if the distribution of k-max$(T_{n,i} : i \in I(P))$ is continuous under P; otherwise, we can still bound the k-FWER by α. Indeed, if $|I(P)| < k$, then there is nothing to prove; otherwise,

$$P\{k \text{ or more false rejections}\} = P\{k\text{-max}(T_{n,i} : i \in I(P)) > c_{n,I(P)}(1-\alpha, k, P)\} \leq \alpha \ ,$$

with equality if the distribution of k-max$(T_{n,i} : i \in I(P))$ is continuous under P. Unfortunately, the test is unavailable as the critical value is in general unknown.

One possible approach is to replace $c_{n,I(P)}(1-\alpha, k, P)$ by $c_{n,I(P)}(1-\alpha, k)$, but this still depends on P through $I(P)$. Since $I(P)$ is unknown, a conservative approach would be to assume all hypotheses are true and replace $c_{n,I(P)}(1-\alpha, k)$ by $c_{n,A}(1-\alpha, k)$, where $A = \{1, \ldots, s\}$.

Unfortunately, in nonparametric problems, the sup in (10) may be formidable or impossible to calculate, and may be way too conservative anyway. Instead, another possibility is to

replace the critical value $c_{n,I(P)}(1-\alpha,k,P)$ by some estimate $\hat{c}_{n,I(P)}(1-\alpha,k)$, which is at least consistent or conservative. In general, suppose $\hat{c}_{n,K}(1-\alpha,k)$ represents an approximation or estimate of the $1-\alpha$ quantile of the distribution of k-max$(T_{n,i}: \ i \in K)$, at least valid when H_i is true for $i \in K$. Bootstrap and subsampling methods offer viable general approaches, and will be used later. Such a single-step approach using the k-max statistic was also discussed in Dudoit et al. (2004). (Rather than formalizing the required conditions for consistency right now, we will later give explicit conditions for more powerful stepdown methods.) A single-step approach would then be to replace K by $A = \{1,\ldots,s\}$. To give concrete representations, we offer two examples.

Example 2.1 (Multivariate Normal Mean) Suppose $(X_1,\ldots,X_s)$ is multivariate normal with unknown mean $\mu = (\mu_1,\ldots,\mu_s)$ and known covariance matrix Σ having (i,j) component $\sigma_{i,j}$. Consider testing $H_i : \mu_i \leq 0$ versus $\mu_i > 0$. Let $T_{n,i} = X_i/\sqrt{\sigma_{i,i}}$, since the test that rejects for large $X_i/\sqrt{\sigma_{i,i}}$ is UMP for testing H_i. For $|K| \geq k$, $c_{n,K}(1-\alpha,k)$ is the $1-\alpha$ quantile of the distribution of k-max$(T_{n,i}: \ i \in K)$ when $\mu = 0$. A single-step approach would reject any $T_{n,i}$ that exceeds $c_{n,A}(1-\alpha,k)$, where $A = \{1,\ldots,s\}$. Since

$$c_{n,A}(1-\alpha,k) \geq c_{n,I(P)}(1-\alpha,k) \geq c_{n,I(P)}(1-\alpha,k,P) \ ,$$

this procedure clearly controls the k-FWER. Of course, it is strictly more powerful than a Bonferroni procedure, since it accounts for the dependence between the test statistics.

In the special case when $k = 1$ and $\sigma_{i,i} = \sigma^2$ is independent of i and $\sigma_{i,j}$ has the product structure $\sigma_{i,j} = \lambda_i\lambda_j$, then Appendix 3 of Hochberg and Tamhane (1987, page 374) reduces the problem of determining the distribution of the maximum of a multivariate normal vector to a univariate integral. For general k or general Σ, one can resort to simulation to approximate the critical values. ■

Outside some parametric models or models where permutation tests apply, exact critical values are usually not available. We now offer a concrete approach based on the bootstrap. The theory of asymptotic control will follow from results for the more powerful stepdown method which we develop later.

Example 2.2 Suppose H_i is concerned with a test of a parameter; that is, H_i is specified by $\{P: \ \theta_i(P) \leq 0\}$ for some real-valued parameter θ_i. Let $\hat{\theta}_{n,i}$ be an estimate of θ_i. Also, let $T_{n,i} = \tau_n\hat{\theta}_{n,i}$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$. The sequence τ_n is introduced for asymptotic purposes so that a limiting distribution for $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ exists. In typical situations, $\tau_n = n^{1/2}$.

The bootstrap method relies on its ability to approximate the joint distribution of $\{\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]: \ i \in K\}$, which we denote by $J_{n,K}(P)$.

For $K \subset \{1,\ldots,s\}$ with $|K| \geq k$, let $L_{n,K}(k,P)$ denote the distribution under P of k-max$(\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]: \ i \in K)$, with corresponding cumulative distribution function $L_{n,K}(x,k,P)$ and α-quantile

$$b_{n,K}(\alpha,k,P) = \inf\{x : L_{n,K}(x,k,P) \geq \alpha\} \ .$$

Let $\hat{Q}_n$ be some estimate of P. For i.i.d. data, $\hat{Q}_n$ is typically taken to be the empirical distribution, or possibly a smoothed version. For time series or data-dependent situations, block bootstrap methods should be employed; see Lahiri (2003). Let $A = \{1, \ldots, s\}$. Then, a nominal $1-\alpha$ level bootstrap joint confidence region for the subset of parameters $\{\theta_i(P) : i \in A\}$ is given by

$$\{(\theta_i : i \in A) : \max_{i \in A} \tau_n[\hat{\theta}_{n,i} - \theta_i] \leq b_{n,A}(1-\alpha, 1, \hat{Q}_n)\} \tag{11}$$

$$= \{(\theta_i : i \in A) : \theta_i \geq \hat{\theta}_{n,i} - \tau_n^{-1} b_{n,A}(1-\alpha, 1, \hat{Q}_n)\} \ .$$

A value of 0 for $\theta_i(P)$ falls outside the region if and only if $\tau_n \hat{\theta}_{n,i} > b_{n,A}(1-\alpha, 1, \hat{Q}_n)$. By the usual duality of confidence sets and hypothesis tests, this suggests the use of the critical value

$$\hat{c}_{n,A}(1-\alpha, 1) = b_{n,A}(1-\alpha, 1, \hat{Q}_n) \ , \tag{12}$$

to control the familywise error rate (i.e. the k-FWER with $k = 1$) at least if the bootstrap is a valid asymptotic approach for joint confidence region construction. Since here, we require control of the k-FWER, we merely replace the max in (11) with the k-max and $b_{n,A}(1-\alpha, 1, \hat{Q}_n)$ with $b_{n,A}(1-\alpha, k, \hat{Q}_n)$. Such a generalized joint confidence region should asymptotically contain all true parameter values except for possibly at most $k-1$ of them, with probability (asymptotically) at least $1-\alpha$. Thus, the bootstrap critical value we use will be

$$\hat{c}_{n,A}(1-\alpha, k) = b_{n,A}(1-\alpha, k, \hat{Q}_n) \ . \tag{13}$$

Asymptotic control of this single-step bootstrap method will follow from later results on the more powerful stepdown bootstrap method of Section 3.1. ■

2.2 Stepdown Methods That Control the k-FWER

Let

$$T_{n,r_1} \geq T_{n,r_2} \geq \cdots \geq T_{n,r_s} \tag{14}$$

denote the observed ordered test statistics, and let H_{r_1}, $H_{r_2}, \ldots, H_{r_s}$ be the corresponding hypotheses.

Stepdown methods begin by first applying a single-step method, but then additional hypotheses may be rejected after this first stage by proceeding in a stepwise fashion, which we now describe. Begin by testing the joint null (intersection) hypothesis $H_{\{1,\ldots,s\}}$ that all hypotheses are true. This hypothesis is rejected if T_{n,r_1} is deemed large, in which case H_{r_1} is rejected. Here, the meaning of large is determined by some critical value $\hat{c}_{n,A}(1-\alpha, k)$, which is designed to offer single-step control when testing the intersection hypothesis H_A with $A = \{1, \ldots, s\}$. If it is not large, accept all hypotheses; otherwise, reject the hypothesis corresponding to the largest test statistic. Once a hypothesis is rejected, the next most significant hypothesis corresponding to the next largest test statistic is considered, and so on. At any stage, one tests appropriate intersection hypotheses H_K. Suppose that critical constants $\hat{c}_{n,K}(1-\alpha, k)$ are available from our statistical tool chest, which we might contemplate for use as a single step procedure for testing H_K. The critical constants $\hat{c}_{n,K}(1-\alpha, k)$ may be fixed or random, but the reader should have in mind that they each could be used as a test of H_K.

Algorithm 2.1 (Generic Stepdown Method For Control of the k-FWER)

1. Let $A_1 = \{1, \ldots, s\}$. If $\max(T_{n,i} : i \in A_1) \leq \hat{c}_{n,A_1}(1-\alpha, k)$, then accept all hypotheses and stop; otherwise, reject any H_i for which $T_{n,i} > \hat{c}_{n,A_1}(1-\alpha, k)$ and continue.

2. Let R_2 be the indices i of hypotheses H_i previously rejected, and let A_2 be the indices of the the remaining hypotheses. If $|R_2| < k$, then stop. Otherwise, let

$$\hat{d}_{n,A_2}(1-\alpha, k) = \max\{\hat{c}_{n,K}(1-\alpha, k) : K = A_2 \cup I,\ I \subset R_2,\ |I| = k-1\} .$$

   Then, reject any $T_{n,i}$ with $i \in A_2$ satisfying $T_{n,i} > \hat{d}_{n,A_2}(1-\alpha, k)$. If there are no further rejections, stop.

$\vdots$

j. Let R_j be the indices i of hypotheses H_i previously rejected, and let A_j be the indices of the remaining hypotheses. Let

$$\hat{d}_{n,A_j}(1-\alpha, k) = \max\{\hat{c}_{n,K}(1-\alpha, k) : K = A_j \cup I,\ I \subset R_j,\ |I| = k-1\} .$$

   Then, reject any $T_{n,i}$ with $i \in A_j$ satisfying $T_{n,i} > \hat{d}_{n,A_j}(1-\alpha, k)$. If there are no further rejections, stop.

$\vdots$

   And so on.

Note that, in the case $k = 1$, once a hypothesis is removed, it no longer enters into the algorithm. However, for $k > 1$, the algorithm becomes slightly more complex. The reason is that, for control of the k-FWER, we must acknowledge that when we consider a set of hypotheses not previously rejected, we may have gotten to that stage in the algorithm by rejecting true null hypotheses, but hopefully at most $k-1$ of them. Since we do not know which of the hypotheses rejected thus far are true or false, we must maximize over subsets including some of those rejected, but at most $k-1$ among the previously rejected ones. Note that, in the case $k = 1$, no previously rejected hypotheses need be considered any further in the determination of whether more hypotheses will be rejected. Thus, the case $k = 1$, as considered in Romano and Wolf (2005a) is particularly simple, especially from a computational point of view. Our main point will be that, if we can control the k-FWER at any stage of the algorithm, then the stepdown test will control the k-FWER.

Remark 2.1 (Modified Generic Stepdown Method For Control of the k-FWER) The following modification of Algorithm 2.1 has the exact same properties in terms of k-FWER control but potentially rejects more false hypotheses: If Algorithm 2.1 rejects at least $k-1$ hypotheses, reject the same hypotheses. Otherwise, reject $H_{r_1}, \ldots, H_{r_{k-1}}$. In other words, the $k-1$ most significant hypotheses are rejected regardless of the data. We do not necessarily

promote this approach, as it can lead to counterintuitive results. Take the case where individual p-values are available and the test statistics are of the form $T_{n,i} = 1 - \hat{p}_i$. For any $k \geq 2$ one would then always reject H_{r_1} even if $\hat{p}_{r_1} = 0.5$, say. On the other hand, for certain applications this modified algorithm may be preferred. ■

In order to prove such an algorithm controls the k-FWER for suitable choice of critical values $\hat{c}_{n,K}(1-\alpha, k)$, we assume monotonicity of the estimated critical values; that is, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha, k) \geq \hat{c}_{n,I(P)}(1-\alpha, k) \ . \tag{15}$$

Ideally, we would also like the following to hold: if $\hat{c}_{n,K}(1-\alpha, k)$ is used to test the intersection hypothesis H_K, then the chance of k or more false rejections is bounded above by α when $K = I(P)$; that is,

$$P\{k\text{-max}(T_{n,i} : \ i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha, k)\} \leq \alpha \ . \tag{16}$$

Under the monotonicity assumption (15), we will show the basic inequality that k-FWER$_P$ is bounded above by left side of (16). This will then show that, if we can construct monotone critical values such that each intersection test controls the k-FWER, then the stepdown procedure controls the k-FWER. Thus, the construction of a stepdown procedure is effectively reduced to construction of single tests, as long as the monotonicity assumption holds. Also, note the monotonicity assumption for the critical values can be made to hold by construction and can be enforced, that is, it does not depend on the unknown P.

Theorem 2.1 *Let P denote the true distribution generating the data. Consider Algorithm 2.1 with critical values $\hat{c}_{n,K}(1-\alpha, k)$ satisfying (15).*

(i) Then,

$$k\text{-}FWER_P \leq P\{k\text{-}\max(T_{n,i} : \ i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha, k)\} \ . \tag{17}$$

(ii) Therefore, if the critical values also satisfy (16), then k-$FWER_P \leq \alpha$.

The monotonicity assumption (15) cannot be removed, as shown in Example 2.1 of Romano and Wolf (2005a), in the case $k = 1$, and an analogous construction works for general k. Fortunately, the general resampling constructions we describe later will inherently satisfy (15).

As a corollary, suppose we consider the nonrandom choice of critical values

$$\hat{c}_{n,K}(1-\alpha, k) = c_{n,K}(1-\alpha, k)$$

defined in (10). Assume the following monotonicity assumption: for $K \supset I(P)$,

$$c_{n,K}(1-\alpha, k) \geq c_{n,I(P)}(1-\alpha, k) \ . \tag{18}$$

The condition (18) can be expected to hold in many situations because the left hand side is based on computing the $1-\alpha$ quantile of the kth largest of $|K|$ variables, while the right hand side is based on the kth largest of $|I(P)| \leq |K|$ variables (though one must be careful and realize that the quantiles are computed under possibly different P, which is why some condition is required).

Corollary 2.1 *Let P denote the true distribution generating the data. Assume $\bigcap_{i=1}^{s} \omega_i$ is not empty.*

(i) Assume (18). Consider Algorithm 2.1 with $\hat{c}_{n,K}(1-\alpha,k) = c_{n,K}(1-\alpha,k)$. Then, k-FWER$_P \leq \alpha$.

(ii) Strong control persists if, in Algorithm 2.1, the critical constants $\hat{c}_{n,K}(1-\alpha,k)$ are replaced by $d_{n,K}(1-\alpha,k)$ which satisfy

$$d_{n,K}(1-\alpha,k) \geq c_{n,K}(1-\alpha,k) \ . \tag{19}$$

(iii) Moreover, the condition (18) may then be removed if the $d_{n,K}(1-\alpha,k)$ satisfy

$$d_{n,K}(1-\alpha,k) \geq d_{n,I(P)}(1-\alpha,k) \tag{20}$$

for any $K \supset I(P)$.

Example 2.3 (Multivariate Normal Mean, continuation of Example 2.1) Recall that $(X_1,\ldots,X_s)$ is multivariate normal with unknown mean $\mu = (\mu_1,\ldots,\mu_s)$ and known covariance matrix Σ having (i,j) component $\sigma_{i,j}$. Consider testing $H_i : \mu_i \leq 0$ versus $\mu_i > 0$. Let $T_{n,i} = X_i/\sqrt{\sigma_{i,i}}$. To apply Corollary 2.1, assume that $|I(P)| \geq k$ or there is nothing to prove. Let $c_{n,K}(1-\alpha,k)$ be the $1-\alpha$ quantile of the distribution of k-max$(T_{n,i} : \ i \in K)$ when $\mu = 0$. Since

$$k\text{-max}(T_{n,i} : \ i \in I) \leq k\text{-max}(T_{n,i} : \ i \in K)$$

whenever $I \subset K$, the monotonicity requirement (18) is satisfied. Moreover, the resulting test procedure rejects at least as many hypotheses as the generalized Holm procedure, as it accounts for the dependence of the test statistics. ■

Example 2.4 (One-way Layout) Suppose for $i = 1,\ldots,s$ and $j = 1,\ldots,n_i$, $X_{i,j} = \mu_i + \epsilon_{i,j}$, where the $\epsilon_{i,j}$ are i.i.d. $N(0,\sigma^2)$; the vector $\mu = (\mu_1,\ldots,\mu_s)$ and σ^2 are unknown. Consider testing $H_i : \mu_i = 0$ against $\mu_i \neq 0$. Let $t_{n,i} = \sqrt{n_i}\bar{X}_{i,\cdot}/S$, where

$$\bar{X}_{i,\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}, \qquad S^2 = \frac{1}{\nu}\sum_{i=1}^{s}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i,\cdot})^2 \ ,$$

and $\nu = \sum_i (n_i - 1)$. Under H_i, $t_{n,i}$ has a t-distribution with ν degrees of freedom. Let $T_{n,i} = |t_{n,i}|$, and let $c_{n,K}(1-\alpha,k)$ denote the $1-\alpha$ quantile of the distribution of k-max$(T_{n,i} : \ i \in K)$ when $\mu = 0$ and $\sigma = 1$. Since, for $|I| \geq k$,

$$k\text{-max}(T_{n,i} : \ i \in I) \leq k\text{-max}(T_{n,i} : \ i \in K) \ ,$$

whenever $I \subset K$, the monotonicity requirement (18) follows. Note that the joint distribution of $(t_{n,1},\ldots,t_{n,s})$ follows an s-variate multivariate t-distribution with ν degrees of freedom; see Hochberg and Tamhane (1987, pp. 374–375). ■

The previous examples are parametric in nature and the null distributions for testing intersection hypotheses do not depend on nuisance parameters. However, we will see that a valid stepdown approach can apply to semiparametric and nonparametric problems if we can construct single step tests of intersection hypotheses whose critical values satisfy the monotonicity requirement. Our main goal will be to apply resampling methods that can implicitly account for the dependence structure of the test statistics. However, we first observe that the fact that the generalized Holm procedure controls the k-FWER follows from Corollary 2.1.

Example 2.5 (Generalized Holm procedure) The stepdown procedure described by (6) with critical values given by (7) controls the k-FWER. This follows from Theorem 2.1 and the fact that, when testing $|K|$ hypotheses, the single-step procedure that rejects any hypothesis for which its corresponding p-value is $\leq k\alpha/|K|$ controls the k-FWER; see Theorem 2.1 (i) of Lehmann and Romano (2005). Note that the critical values $k\alpha/|K|$ are monotone in $|K|$. ■

Remark 2.2 In general, the critical values used in Corollary 2.1(i) are the smallest constants possible without violating the k-FWER. As a simple example, suppose X_i, $i = 1, \ldots, s$, are independent $N(\theta_i, 1)$, with the θ_i varying freely. The null hypothesis H_i specifies $\theta_i \leq 0$ and $T_{n,i} = X_i$. Then, $c_{n,K}(1-\alpha, k)$ is the $1-\alpha$ quantile of k-max$(Z_1, \ldots, Z_{|K|})$, where the Z_i are i.i.d. $N(0,1)$. Suppose c is a constant and $c < c_{n,K}(1-\alpha, k)$ for some subset K with $|K \cap I(P)| \geq k$. As $\theta_i \to \infty$ for $i \notin K$ and $\theta_i = 0$ for $i \in K$, the probability of k or more false rejections tends to

$$P\{k\text{-max}(X_i : i \in K) > c\} > P\{k\text{-max}(X_i : i \in K) > c_{n,K}(1-\alpha, k)\} = \alpha \ .$$

Thus, the sup over P of the probability (under P) that Algorithm 2.1 rejects any $i \in I(P)$ is equal to α. It then follows that the critical values cannot be made smaller, in hopes of increasing the ability to detect false hypotheses, without violating the strong control of the k-FWER. However, the above only applies to nonrandom critical values and does not negate the possibility that critical values can be estimated, and therefore be random. That is, if we replace $c_{n,K}(1-\alpha, k)$ by some estimate $\hat{c}_{n,K}(1-\alpha, k)$, it can sometimes be smaller than $c_{n,K}(1-\alpha, k)$ as long as it is not with probability one. Of course, it is typically the case that critical values need to be estimated, such as by permutation tests, resampling, bootstrap and subsampling methods, and these will be considered in the later sections. ■

In the examples considered so far, the application of the Generic Stepdown Method was not highly computational because the critical values essentially only depended on the number of hypotheses being tested at any stage. When this is not the case, the procedure becomes more computational. However, we will also consider the following more streamlined algorithm. The basic idea is that at any stage, when testing whether or not to include further rejections, we need only look at the hypotheses not previously rejected together with the $k-1$ hypotheses that are least significant among those previously rejected. So, we avoid maximizing over all subsets of size $k-1$ of previously rejected hypotheses and just look at the most "recent" $k-1$ rejections. The arguments for such a procedure will be asymptotic. The algorithm looks like this.

Algorithm 2.2 (Streamlined Stepdown Method For Control of the k-FWER)

1. Let $A_1 = \{1, \ldots, s\}$. If $\max(T_{n,i} : i \in A_1) \leq \hat{c}_{n,A_1}(1-\alpha, k)$, then accept all hypotheses and stop; otherwise, reject any H_i for which $T_{n,i} > \hat{c}_{n,A_1}(1-\alpha, k)$ and continue.

2. Let R_2 be the indices i of hypotheses H_i previously rejected, and let A_2 be the indices of the remaining hypotheses. If $R_2 < k$, then stop. Otherwise, let K be the union of A_2 together with the $k-1$ least significant hypotheses among those previously rejected, so

$$K = \{r_{(|R_2|-k+2)}, r_{(|R_2|-k+3)}, \ldots, r_{(s)}\} .$$

   Set

$$\tilde{d}_{n,A_2}(1-\alpha, k) = \hat{c}_{n,K}(1-\alpha, k) .$$

   Then, reject any $T_{n,i}$ with $i \in A_2$ satisfying $T_{n,i} > \tilde{d}_{n,A_2}(1-\alpha, k)$. If there are no further rejections, stop.

⋮

j. Let R_j be the indices i of hypotheses H_i previously rejected, and let A_j be the indices of the remaining hypotheses. Let K be the union of A_j together with the $k-1$ least significant hypotheses among those previously rejected, so

$$K = \{r_{(|R_j|-k+2)}, r_{(|R_j|-k+1)}, \ldots, r_{(s)}\} .$$

   Let

$$\tilde{d}_{n,A_j}(1-\alpha, k) = \hat{c}_{n,K}(1-\alpha, k) .$$

   Then, reject any $T_{n,i}$ with $i \in A_j$ satisfying $T_{n,i} > \tilde{d}_{n,A_j}(1-\alpha, k)$. If there are no further rejections, stop.

⋮

   And so on.

2.3 Permutation and Randomization Tests

We now show how Theorem 2.1 can be applied to permutation and randomization tests. First, we review a general construction of a randomization test in the context of a single test, because the key result of Theorem 2.1 is that the general problem of constructing valid stepdown tests can be reduced to the construction of tests of intersection hypotheses, as long as we can verify the monotonicity requirement. Our setup is framed in terms of a population model, but similar results are possible in terms of a randomization model (as in Section 3.1.7 of Westfall and Young, 1993).

Based on data X taking values in a sample space $\mathcal{X}$, it is desired to test the null hypothesis H that the underlying probability law P generating X belongs to a certain family

ω of distributions. Let $\mathbf{G}$ be a finite group of transformations g of $\mathcal{X}$ onto itself. The following assumption, which we will call the *randomization hypothesis*, allows for a general test construction.

The Randomization Hypothesis The null hypothesis implies that the distribution of X is invariant under the transformations in $\mathbf{G}$; that is, for every g in $\mathbf{G}$, gX and X have the same distribution whenever X has distribution P in ω.

As an example, consider testing the equality of distributions based on two independent samples $(Y_1, \ldots, Y_m)$ and $(Z_1, \ldots, Z_n)$. Under the null hypothesis that the samples are generated from the same probability law, the observations can be permuted or assigned at random to either of the two groups, and the distribution of the permuted samples is the same as the distribution of the original samples. In this example, and more generally when the randomization hypothesis holds, the following construction of a randomization test applies.

Let $T(X)$ be any real-valued test statistic for testing H. Suppose the group $\mathbf{G}$ has M elements. Given $X = x$, let

$$T^{(1)}(x) \leq T^{(2)}(x) \leq \cdots \leq T^{(M)}(x)$$

be the values of $T(gx)$ as g varies in $\mathbf{G}$, ordered from smallest to largest. Fix a nominal level α, $0 < \alpha < 1$, and let m be defined by

$$m = M - \lfloor M\alpha \rfloor \ , \tag{21}$$

where $\lfloor M\alpha \rfloor$ denotes the largest integer less than or equal to $M\alpha$. Let $M^+(x)$ and $M^0(x)$ be the number of values $T^{(j)}(x)$ $(j = 1, \ldots, M)$ which are greater than $T^{(m)}(x)$ and equal to $T^{(m)}(x)$, respectively. Set

$$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)} \ .$$

Define the randomization test function $\phi(X)$ to be equal to 1, $a(X)$, or 0 according to whether $T(X) > T^{(m)}(X)$, $T(X) = T^{(m)}(X)$, or $T(X) < T^{(m)}(X)$, respectively.

Under the randomization hypothesis, Hoeffding (1952) shows this construction produces a test that is exact level α, and this result is true for *any* choice of test statistic T. Note that this test is possibly a randomized test if $M\alpha$ is not an integer of there are ties in the ordered values. Alternatively, if one prefers not to randomize, the slightly conservative but *nonrandomized* test that rejects if $T(X) > T^m(X)$ is level α.

In general, one can define a p-value $\hat{p}$ of a randomization test by

$$\hat{p} = \frac{1}{M} \sum_g I\{T(gX) \geq T(X)\} \ . \tag{22}$$

It is easily shown that $\hat{p}$ satisfies, under the null hypothesis,

$$P\{\hat{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ . \tag{23}$$

Therefore, the *nonrandomized* test that rejects when $\hat{p} \leq \alpha$ is level α.

We now return to the multiple testing problem. Assume $\mathbf{G}_K$ is a group of transformations for which the randomization hypothesis holds for H_K. Then, if we wish to control the k-FWER, we can apply the above construction to test the single intersection hypothesis H_K based on the test statistic

$$T_{n,K} = k\text{-max}(T_{n,i} : \ i \in K) \tag{24}$$

and reject H_K when

$$T_{n,K}(X) > T_{n,K}^{(|\mathbf{G}_K| - \lfloor |\mathbf{G}_K|\alpha \rfloor)}(X) \ .$$

If it is also the case that $\mathbf{G}_K = \mathbf{G}$, so that the same $\mathbf{G}$ applies to all intersection hypotheses, then we can verify the monotonicity assumption for the critical values. Set $m_\alpha = |\mathbf{G}| - \lfloor |\mathbf{G}|\alpha \rfloor$. Then, for any $g \in \mathbf{G}$ and $I \subset K$,

$$k\text{-max}(T_{n,i}(gX) : \ i \in K) \geq k\text{-max}(T_{n,i}(gX) : \ i \in I) \ , \tag{25}$$

and so as g varies, the m_αth largest value of the left side of (25) is at least as large as the m_αth largest value of the right side.

Consequently, the critical values

$$\hat{c}_{n,K}(1-\alpha, k) = T_{n,K}^{(m_\alpha)} \ , \tag{26}$$

satisfy the monotonicity requirement of Theorem 2.1. Moreover, by the general randomization construction of a single test, the test that rejects H_K when $T_{n,K} \geq T_{n,K}^{(m_\alpha)}$ is level α. Therefore, the following is true.

Corollary 2.2 *Suppose the randomization hypothesis holds for a group* $\mathbf{G}$ *when testing any intersection hypothesis* H_K*. Then, the stepdown method with critical values given by (26) controls the k-FWER at level* α*.*

Remark 2.3 Because $\mathbf{G}$ may be large, one may resort to a stochastic approximation to construct the randomization test, by randomly sampling transformations g from $\mathbf{G}$. The results are valid in this case; see Romano and Wolf (2005) who considered the case $k = 1$, but the results generalize.

In the above corollary, we have worked with the randomization construction using nonrandomized tests. A similar result would hold if we permit randomization. ■

Example 2.6 (Two Sample Problem With k Variables) Suppose $Y_1, \cdots, Y_{n_Y}$ is a sample of n_Y independent observations from a probability distribution P_Y and $Z_1, \cdots, Z_{n_Z}$ is a sample of n_Z observations from P_Z. Here, P_Y and P_Z are probability distributions on $\mathbb{R}^s$, with ith components denoted $P_{Y,i}$ and $P_{Z,i}$, respectively. The hypothesis H_j asserts $P_{Y,i} = P_{Z,i}$ and we wish to test these k hypotheses based on $X = (Y_1, \cdots, Y_{n_Y}, Z_1, \cdots, Z_{n_Z})$. Also, let $Y_{j,i}$ denote the ith component of Y_j and $Z_{j,i}$ denote the ith component of Z_j. As in Troendle (1995), we assume a semiparametric model. In particular, assume P_Y and P_Z are governed by a family of probability distributions Q_θ indexed by $\theta = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s$ (and assumed identifiable), so that P_Y has law $Q(\theta_Y)$ and P_Z has law $Q(\theta_Z)$. For concreteness, one may think of θ as being

the mean vector, though this assumption is not necessary. Now, H_i can be viewed as testing $\theta_{Y,i} = \theta_{Z,i}$. Note that the randomization construction does not need to assume knowledge of the form of Q (just as a single two-sample permutation test in a shift model does not need to know the form of the underlying distribution under the null hypothesis).

Let $n = n_Y + n_Z$, and for $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$, let $gx \in \mathbb{R}^n$ be defined by $(x_{\pi(1)}, \cdots, x_{\pi(n)})$, where $(\pi(1), \cdots, \pi(n))$ is a permutation of $(1, 2, \cdots, n)$. Let $\mathbf{G}$ be the collection of all such g so that $M = n!$. Under the hypothesis $P_Y = P_Z$, gX and X have the same distribution for any g in $\mathbf{G}$.

Unfortunately, this $\mathbf{G}$ does not apply to any subset K of the hypotheses, because gX and X need not have the same distribution if only a subcollection of the hypotheses are true. However, we just need a slight generalization to cover the example. Suppose that the test statistic $T_{n,i}$ used to test H_i only depends on the ith components of the observations, namely $Y_{j,i}$, $j = 1, \ldots, n_Y$ and $Z_{j,i}$, $j = 1, \ldots, n_Z$; this is a weak assumption indeed. In fact, let X_K be the data set consisting of the the components $Y_{j,i}$ and $Z_{j,i}$ as i varies only in K. The simple but important point here is that, for this reduced data set, the randomization hypothesis holds. Specifically, under the null hypothesis $\theta_{Y,i} = \theta_{Z,i}$ for $i \in K$, X_K and gX_K have the same distribution (though X and gX need not). Also, for any $g \in \mathbf{G}$, $T_{n,i}(gX)$ and $T_{n,i}(X)$ have the same distribution under H_i, and similarly for any $K \subset \{1, \ldots, s\}$, $T_{n,K}(gX)$ and $T_{n,K}(X)$ have the same distribution under H_K.

Then, because the same $\mathbf{G}$ applies in this manner for all K, the critical values from the randomization test are monotone, just as in (25). Moreover, each intersection hypothesis can be tested by an exact level α randomization test (since inference for H_K is based only on X_K). Therefore, essentially the same argument leading to Corollary 2.2 applies. In particular, even if we need to resort to approximate randomization tests at each stage, but as long as we sample the same set of g_j from $\mathbf{G}$, the resulting procedure retains its finite sample property of controlling the k-FWER. In contrast, Troendle (1995), discussing the special case of $k = 1$, concludes asymptotic control only. For general k, Korn et al. (2004) discuss finite sample control of the k-FWER in the setting of this example ■

Example 2.7 (Semiparametric version of Example 2.3) Suppose, for $j = 1, \ldots, n$, X_j are i.i.d. s-variate with $X_j = (X_{j,1}, \ldots, X_{j,s})$. It is assumed $X_j = \mu + \epsilon_j$, where $\mu = (\mu_1, \ldots, \mu_s)$ and the ϵ_j are i.i.d. random vectors with s-variate distribution F. The distribution of F is unknown, but greatly weakening the assumption of multivariate normality, it is assumed that the distribution of F is symmetric in the sense that the distribution of ϵ_j is the same as that of $-\epsilon_j$. Consider testing $H_0 : \mu_i = 0$ against $\mu_i \neq 0$. Let $t_{n,i} = \sqrt{n}\bar{X}_{n,i}/S_{n,i}$, where

$$\bar{X}_{n,i} = \frac{1}{n}\sum_{j=1}^{n} X_{j,i}, \qquad S_{n,i}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_{j,i} - \bar{X}_{n,i})^2 \ ;$$

also, set $T_{n,i} = |t_{n,i}|$. To test the intersection hypothesis H_K, consider the group $\mathbf{G}_K$ of 2^n transformations of the form

$$(X_1, \ldots, X_n) \to (\delta_1 X_1, \ldots, \delta_n X_n) \ ,$$

where the δ_i are either 1 or -1. These transformations apply to any K, but as in the previous example, the randomization hypothesis strictly speaking does not hold for testing H_K. However, as in the previous example, $T_{n,K}(X)$ and $T_{n,K}(gX)$ have the same distribution under H_K and the argument leading to Corollary 2.2 applies to yield exact finite sample control of the k-FWER. ■

Remark 2.4 It is interesting to study the behavior of randomization and permutation procedures if the model is such that the randomization hypothesis does not hold. For example, in Example 2.7, we may be interested in testing $H_j : \mu_i = 0$ even if $X_{j,i}$ is not assumed to have a symmetric distribution. Then, the randomization test construction of this section fails because the randomization hypothesis need not hold. However, since the randomization procedure has monotone critical values (as this is only a property of how the data is used), Theorem 2.1(i) applies. Therefore, one can again reduce the problem of studying control of the k-FWER to that of controlling the level of a single intersection hypothesis. But the problem of controlling the level of a single test when the randomization hypothesis fails is studied in Romano (1990) and so similar methods can be used here, with the hope of at least proving asymptotic control. Alternatively, the more general resampling approaches of Section 3 can be employed; the comparison of randomization and bootstrap tests is studied in Romano (1989) and it is shown they are often quite close, at least when the randomization hypothesis holds. ■

Example 2.8 (Comparison of Multiple Treatments with a Control) Consider the one-way anova model. We are given $s+1$ independent samples, with the ith sample having n_i i.i.d. observations $X_{i,j}$, $j = 1, \ldots, n_i$. Suppose $X_{i,j}$ has distribution P_i. The problem is to test the hypotheses of s treatments with a control; that is, $H_i : P_i = P_{s+1}$. (Alternatively, we can test all pairs of distributions, but the issues are much the same, so we illustrate them with the slightly easier setup.) Under the joint null hypothesis, we can randomly assign all $n = \sum_{i=1}^{s+1} n_i$ observations to any of the groups; that is, the group $\mathbf{G}$ consists of all permutations of the data. However, if only a subset of the hypotheses are true, this group is not valid. A simple remedy is to permute only within subsets; that is, to test any subset hypothesis H_K, only consider those permutations that permute observations within the samples $X_{i,j}$ with $i \in K$ and the sample $X_{s+1,j}$. Therefore, one computes a critical value $\hat{c}_{n,K}(1-\alpha, k)$ by the randomization test with the group $\mathbf{G}_K$ of permutations within samples $i \in K$ and $i = s+1$. Unfortunately, this does not lead to monotonicity of critical values, and the previous results do not apply. But, we can apply the generalized closure method of the appendix, if one is willing to compute critical values for all subset hypotheses. On the other hand, this can be computationally prohibitive. Such issues are raised by Petrondas and Gabriel (1983) (although the problem was not framed in terms of a monotonicity requirement). Fortunately, the lack of monotonicity of critical values is only a concern if strict finite sample control is required; otherwise, computationally quicker bootstrap methods described in the next section apply to yield asymptotic control. ■

3 Asymptotic Results on k-FWER Control

The main goal of this section is to show how Theorem 2.1 can be used to construct stepdown procedures that asymptotically control the k-FWER under very weak assumptions. The use of resampling techniques will be a key ingredient. The assumptions are identical to the weakest assumptions available for the contruction of asymptotically valid tests of a single hypothesis, which are used in many resampling schemes, and so one cannot expect to improve them without improving the now well-developed theory of resampling methods for testing a single hypothesis. The methods constructed will be based in Algorithm 2.1, and so many tests are constructed in a stepwise fashion. However, a key feature is that the methods will only require *one* set of resamples for all of the tests, whether they are bootstrap samples or subsamples.

In order to accomplish this, we will consider resampling schemes that do *not* obey the null hypothesis constraints. Hypothesis test constructions that do obey the constraints imposed the null hypothesis, as discussed in Beran (1986) and Romano (1988), are based on the idea that the critical value should be obtained under the null hypothesis and so the resampling scheme should reflect the constraints of the null hypothesis. This idea is even advocated as a principle in Hall and Wilson (1991), and it is enforced throughout Westfall and Young (1993). While appealing, it is by no means the only approach toward inference in hypothesis testing. Indeed, the well-known explicit duality between tests and confidence intervals means that if you can construct good or valid confidence intervals, then you can construct good or valid tests, and conversely. For example, by resampling from the empirical distribution to construct a confidence interval for a single parameter, very desirable intervals can be constructed, which would then translate into desirable tests. The same holds for simultaneous confidence sets and multiple tests.

That is not to say that the approach of obeying the null constraints is less appealing. It is, however, often more difficult to apply, and it is unlikely that one resampling scheme obeying the constraints of all hypotheses would work in general in the multiple testing framework. An alternative approach would be to resample from a different distribution at each step, obeying the constraints of the null hypotheses imposed at each step. This approach would probably succeed in a fair amount of generality, but even so, two problems would remain. First, it may be difficult to determine the appropriate resampling sheme for testing each subset hypothesis. Second, even if one knew how to resample at each stage, there is increased computation. Our approach avoids these complications. In some problems, the subset pivotality condition of Westfall and Young (1993) holds, and so the same null distribution can be used at each step. However, this condition does not hold in general, as the following example shows.

Example 3.1 (Testing Correlations) Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$, with $X_i = (X_{i,1}, \ldots, X_{i,d})$. Assume $E|X_{i,j}|^2 < \infty$ and $Var(X_{i,j}) > 0$. Then, the correlation between $X_{1,i}$ and $X_{1,j}$, namely $\rho_{i,j}$ is well-defined. Let $H_{i,j}$ denote the hypothesis that $\rho_{i,j} = 0$, so that the multiple testing problem consists in testing all $s = \binom{d}{2}$ pairwise correlations. Also let $T_{n,i,j}$ denote the ordinary sample correlation between variables i and j. (Note that we are indexing hypotheses and test statistics now by 2 indices i and j.) As noted by Westfall and Young

(1993), Example 2.2, p. 43, subset pivotality fails here. For example, using results of Aitken (1969) and Aitken (1971), if $d = s = 3$, $H_{1,2}$ and $H_{1,3}$ are true but $H_{2,3}$ is false, the joint limiting distribution of $n^{1/2}(T_{n,1,2}, T_{n,1,3})$ is bivariate normal with means zero, variances one, and correlation $\rho_{2,3}$. As acknowledged by Westfall and Young (1993), their methods fail to address this problem (even asymptotically). ■

We shall consider two concrete applications of Theorem 2.1, the first based on the bootstrap and the second based on subsampling. The symbols $\xrightarrow{L}$ and $\xrightarrow{P}$ will denote convergence in law (or distribution) and convergence in probability, respectively.

3.1 A Bootstrap Construction

We now apply Theorem 2.1 to develop an asymptotically valid approach based on the bootstrap, but specializing to the case where H_i is concerned with a test of a parameter. Suppose hypothesis H_i is specified by $\{P : \theta_i(P) \leq 0\}$ for some real-valued parameter θ_i. Implicitly, the alternatives are one-sided, but the two-sided case can be similarly handled. Suppose $\hat{\theta}_{n,i}$ is an estimate of θ_i. Also, let $T_{n,i} = \tau_n \hat{\theta}_{n,i}$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$. The sequence τ_n is introduced for asymptotic purposes so that a limiting distribution for $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ exists. In typical situations, $\tau_n = n^{1/2}$. (It is possible to let τ_n vary with the hypothesis i. Extensions to cases where τ_n depends on P are also possible, using ideas in Politis et al., 1999, Chapter8.)

The bootstrap method relies on its ability to approximate the joint distribution of $\{\tau_n[\hat{\theta}_{n,i} - \theta_i(P)] : i \in K\}$, which we denote by $J_{n,K}(P)$.

For $K \subset \{1, \ldots, s\}$ with $|K| \geq k$, let $L_{n,K}(k, P)$ denote the distribution under P of k-max$(\tau_n[\hat{\theta}_{n,i} - \theta_i(P)] : i \in K)$, with corresponding cumulative distribution function $L_{n,K}(x, k, P)$ and α-quantile

$$b_{n,K}(\alpha, k, P) = \inf\{x : L_{n,K}(x, k, P) \geq \alpha\} \ .$$

We will assume the normalized estimates satisfy the following.

Assumption B1(i) $J_{n,I(P)}(P) \xrightarrow{L} J_{I(P)}(P)$, a nondegenerate limit law.

Assumption B1(i) implies $L_{n,I(P)}(k, P)$ has a limiting distribution $L_{I(P)}(k, P)$. Indeed, the k-max function is a continuous function and the continuous mapping theorem applies; see Lemma B.1. We will assume this limit law satisfies the following mild assumption.

Assumption B1(ii) $L_{I(P)}(\cdot, k, P)$ is continuous and strictly increasing on its support.

Under Assumption B1, it follows that

$$b_{n,I(P)}(1 - \alpha, k, P) \to b_{I(P)}(1 - \alpha, k, P) \ , \tag{27}$$

where $b_{I(P)}(\alpha, k, P)$ is the α-quantile of the limiting distribution $L_{I(P)}(k, P)$.

Let $\hat{Q}_n$ be some estimate of P. For i.i.d. data, $\hat{Q}_n$ is typically taken to be the empirical distribution, or possibly a smoothed version. For time series or data-dependent situations, block bootstrap methods should be employed; see Lahiri (2003). Then, a nominal $1 - \alpha$ level bootstrap joint confidence region for the subset of parameters $\{\theta_i(P) : i \in K\}$ is given by

$$\{(\theta_i : i \in K) : \max(\tau_n[\hat{\theta}_{n,i} - \theta_i] : i \in K) \leq b_{n,K}(1 - \alpha, 1, \hat{Q}_n)\} \tag{28}$$

$$= \{(\theta_i : \ i \in K) : \theta_i \geq \hat{\theta}_{n,i} - \tau_n^{-1} b_{n,K}(1-\alpha, 1, \hat{Q}_n)\} \ .$$

So a value of 0 for $\theta_i(P)$ falls outside the region if and only if $\tau_n \hat{\theta}_{n,i} > b_{n,K}(1-\alpha, 1, \hat{Q}_n)$. By the usual duality of confidence sets and hypothesis tests, this suggests the use of the critical value

$$\hat{c}_{n,K}(1-\alpha, 1) = b_{n,K}(1-\alpha, 1, \hat{Q}_n) \ , \tag{29}$$

to control the familywise error rate (i.e. the k-FWER with $k = 1$) at least if the bootstrap is a valid asymptotic approach for joint confidence region construction. Since here, we require control of the k-FWER, we merely replace the max in (28) with the k-max and $b_{n,K}(1-\alpha, 1, \hat{Q}_n)$ with $b_{n,K}(1-\alpha, k, \hat{Q}_n)$. Such a generalized joint confidence region should asymptotically contain all true parameter values except for possibly at most $k-1$ of them, with probability (asymptotically) at least $1-\alpha$. Thus, the bootstrap critical value we use will be

$$\hat{c}_{n,K}(1-\alpha, k) = b_{n,K}(1-\alpha, k, \hat{Q}_n) \ . \tag{30}$$

Note that, regardless of asymptotic behavior, the monotonicity assumption (15) is always satisfied for the choice (30). Indeed, for any Q and if $I \subset K$, $b_{n,I}(1-\alpha, k, Q)$ is the $1-\alpha$ quantile under Q of the k-max of $|I|$ variables, while $b_{n,K}(1-\alpha, k, Q)$ is the $1-\alpha$ quantile of the k-max of these same $|I|$ variables together with additional $|K| - |I|$ variables.

This simple observation together with Theorem 2.1 immediately implies the following.

Corollary 3.1 *Under the setup and notation of this subsection, consider Algorithm 2.1 with critical values given by (30). Then,*

$$k\text{-}FWER_P \leq P\{k\text{-}\max(T_{n,i} : \ i \in I(P)) > b_{n,I(P)}(1-\alpha, k, \hat{Q}_n)\} \ . \tag{31}$$

Therefore, in order to conclude $\limsup_n k\text{-FWER}_P \leq \alpha$, it is now only necessary to study the asymptotic behavior of $b_{n,K}(1-\alpha, k, \hat{Q}_n)$ in the case $K = I(P)$. For this, we further assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that $\theta_i(P) \leq 0$ for all $i \in I(P)$; that is, we assume the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ for $i \in I(P)$. Specifically, consider the following.

Assumption B2 For any metric ρ metrizing weak convergence on $\mathbb{R}^{|I(P)|}$,

$$\rho\left(J_{n,I(P)}(P), J_{n,I(P)}(\hat{Q}_n)\right) \xrightarrow{P} 0 \ .$$

In addition, we shall also need the following strengthened version.

Assumption B3 Assumptions B1 and B2 hold when $I(P)$ is replaced by $\{1, \ldots, s\}$.

Theorem 3.1 *Fix P satisfying Assumption B1. Let $\hat{Q}_n$ be an estimate of P satisfying B2. Consider the stepdown method in Algorithm 2.1 with $\hat{c}_{n,K}(1-\alpha, k)$ replaced by $b_{n,K}(1-\alpha, k, \hat{Q}_n)$.*

(i) Then, $\limsup_n \ k\text{-}FWER_P \leq \alpha$.

(ii) Suppose the stronger Assumption B3 holds. If P is such that $i \notin I(P)$, i.e. H_i is false and $\theta_i(P) > 0$, then the probability that the stepdown method rejects H_i tends to one.

Example 3.2 (Continuation of Example 3.1) The analysis of sample correlations is a special case of the smooth function model studied in Hall (1992), and the bootstrap approach is valid for such models. ■

Remark 3.1 The $\limsup_n$ in part (i) of Theorem 3.1 can be replaced by a $\lim_n$ if the stronger Assumption B3 holds. Furthermore, if the limit law $J_{\{1,\ldots,s\}}(P)$ has a positive density everywhere, then the weak inequality can be replaced by an equality if and only if there exists at least one $\theta_i = 0$ and no $\theta_i < 0$. On the other hand, if there exists at least one $\theta_i < 0$, then the weak inequality becomes a strict inequality. (The situation is quite analogous to the single testing problem of testing whether a normal mean θ with known variance 1 is ≤ 0 versus > 0; here, the actual rejection probability is strictly less than α if $\theta < 0$.) Therefore, it is in general not possible to have a limiting k-FWER of exactly equal to α. Romano and Wolf (2005b) show this in their Theorem 3.1 for the special case $k = 1$, but the argument generalizes to arbitrary $k \geq 1$. ■

Remark 3.2 The main reason why the bootstrap works here can be traced to the simple result Theorem 2.1. The bootstrap approach, by resampling from a fixed distribution, generates monotone critical values. Therefore, since we know how to construct valid bootstrap tests for each intersection hypothesis, this leads to valid multiple tests. But we learn more. If the bootstrap approximation to the distribution of the k-max is valid to order $O(\epsilon_n)$ in the sense that the probability on the right side of (31) is equal to $\alpha + O(\epsilon_n)$, then we also can deduce $\limsup_n k\text{-FWER}_P \leq \alpha + O(\epsilon_n)$. In other works, if a bootstrap method has good performance for the construction of a single sampling distribution of a real-valued statistic, then this translates into good performance of the bootstrap for constructing stepdown multiple tests. ■

Remark 3.3 The bootstrap can also give dramatic finite-sample gains by accommodating non-normalities, even when the test statistics are independent; for example, see Westfall and Young (1993, page 162) and Westfall and Wolfinger (1997). ■

Remark 3.4 Typically, the asymptotic behavior of a test procedure when P is true will satisfy that it is consistent in the sense that all false hypotheses will be rejected with probability tending to one (as is the case under Theorem 3.1). However, one can also study the behavior of procedures against contiguous alternatives so that not all false hypotheses are rejected with probability tending to one under such sequences. But, of course, if alternative hypotheses are in some sense close to their respective null hypotheses, then the procedures will typically reject even fewer hypotheses, and so the limiting probability of k or more false rejections under a sequence of contiguous alternatives should then be bounded above by α. ■

Remark 3.5 In addition to constructing procedures that control the k-FWER, one typically would like to choose test statistics that lead to procedures that are balanced in the sense that all tests have about the same power. As argued by Beran (1988a), Tu and Zhou (2000), and Rogers and Hsu (2001), balance can be desirable. Alternatively, lack of balance may be desirable so that certain tests are given more weight; see Westfall and Young (1993, page 162) and

Westfall and Wolfinger (1997). While the goal of this paper has been the evaluation of significance while maintaining strong control based on given test statistics, achieving balance is best handled by appropriate choice of test statistics. For example, transforming test statistics to p-values and then using the negative p-values as the basic statistics will lead to better balance. Quite generally, Beran's prepivoting transformation can lead to balance; see Beran (1988a, 1988b). The assumptions of our theorem must then hold for the transformed test statistics. Alternatively, balance can sometimes be achieved by studentization. The construction developed in this subsection can be extended to the case of studentized test statistics. Romano and Wolf (2005b) detail the use of studentized statistics for the special case of FWER control. The generalization to general k-FWER control is straightforward and left to the reader. ■

We now briefly consider the two-sided case. Suppose H_i specifies $\theta_i(P) = 0$ against the alternative $\theta_i(P) \neq 0$. Let $L'_{n,K}(k, P)$ denote the distribution under P of k-max$(\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| : i \in K)$ with corresponding distribution function $L'_{n,K}(x, k, P)$ and α-quantile

$$b'_{n,K}(\alpha, k, P) = \inf\{x : L'_{n,K}(x, k, P) \geq \alpha\} .$$

Accordingly, $L'_K(k, P)$ denotes the limiting distribution of $L'_{n,K}(k, P)$. Finally, let $T'_{n,i} = \tau_n|\hat{\theta}_{n,i}|$.

Theorem 3.2 *Fix P satisfying Assumption B1, but with $L_{I(P)}(k, P)$ in B1(ii) replaced by $L'_{I(P)}(k, P)$. Let $\hat{Q}_n$ be an estimate of P satisfying Assumption B2. Consider the stepdown method in Algorithm 2.1 using the test statistics $T'_{n,i}$ and with $\hat{c}_{n,K}(1-\alpha, k)$ replaced by $b'_{n,K}(1-\alpha, k, \hat{Q}_n)$.*

(i) Then, $\limsup_n k$-FWER$_P \leq \alpha$.

(ii) Suppose the stronger Assumption B3 holds, but with $L_{\{1,\ldots,s\}}(k, P)$ in B1(ii) replaced by $L'_{\{1,\ldots,s\}}(k, P)$. If P is such that $i \notin I(P)$, i.e. H_i is false and $\theta_i(P) \neq 0$, then the probability that the stepdown method rejects H_i tends to one.

(iii) Moreover, if the above algorithm rejects H_i and it is declared that $\theta_i > 0$ when $\hat{\theta}_{n,i} > 0$, the the probability of making a Type 3 error (i.e. of declaring $\theta_i(P)$ positive when it is negative or declaring it negative when it is positive) tends to 0.

An alternative approach to the two-sided case is to balance the tails of the bootstrap distribution of the original estimates (without the absolute values) separately. An analogous result would hold. The comparison of these approaches in the case of a single test is made in Hall (1992).

The result (iii) shows that the directional error is asymptotically negligible. It would be more interesting to obtain both finite sample results, as well as studying the behavior of the directional error under contiguous alternatives so that the problem is no longer asymptotically degenerate; future work will consider these problems. For references to the literature on controlling the directional error as well as some finite sample results, see Finner (1999).

So far, the bootstrap construction has been based on the generic Algorithm 2.1. The following theorem shows that asymptotic control of the k-FWER is also achieved by the computationally less expensive streamlined Algorithm 2.2. For brevity we only focus on the one-sided case, that is, the setting of Theorem 3.1; the result for the two-sided case is very similar.

Theorem 3.3 *Fix P satisfying Assumption B3. Consider the stepdown method in Algorithm 2.2 with $\hat{c}_{n,K}(1-\alpha,k)$ replaced by $b_{n,K}(1-\alpha,\hat{Q}_n,k)$.*

(i) If P is such that $i \notin I(P)$, i.e. H_i is false and $\theta_i(P) > 0$, then the probability that the stepdown method rejects H_i tends to one.

(ii) $\limsup_n$ *k-FWER$_P \leq \alpha$.*

Remark 3.6 The proof of both Theorems 3.1 and 3.3 rely on asymptotic arguments. Nevertheless, there exists an important difference. In essence, Theorem 3.1 rests on the fact the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ for $i \in I(P)$. So does Theorem 3.3, but it also uses the additional fact that, with probability tending to one, all false hypotheses are rejected before any true hypothesis comes under scrutiny. This difference has a number of (related) implications one should keep in mind.

First, the method based on the generic Algorithm 2.1 is more conservative than the one based on the streamlined Algorithm 2.2: the latter will reject all the hypotheses rejected by the former and potentially some further hypotheses.

Second, if instead of the estimated critical values $b_{n,K}(1-\alpha,k,\hat{Q}_n)$ the exact critical values $b_{n,K}(1-\alpha,k,P)$ could be used in place of $\hat{c}_{n,K}(1-\alpha,k)$, then Algorithm 2.1 would provide finite sample control of the k-FWER while Algorithm 2.2 would not.

Third, the bootstrap construction based on Algorithm 2.1 provides asymptotic control of k-FWER in the case of contiguous alternatives while the construction based on Algorithm 2.1 may not. (The reason is that when alternatives are contiguous, the corresponding hypotheses will not be rejected with probability tending to one.) ■

Remark 3.7 (Operative Method) The previous remark provides some motivation to base the bootstrap construction on the more conservative generic Algorithm 2.1. On the other hand, its computational burden can be very high. To compute the critical value $\tilde{d}_{n,A_j}(1-\alpha,k)$ in the jth step, one has to evaluate $N_j = \binom{R_j}{k-1}$ quantiles $\hat{c}_{n,K}(1-\alpha,k)$ in order to then take the largest one of those. Depending on R_j and k, this number N_j may be very large.

Therefore, we now suggest a operative method that retains some of the desirable properties of Algorithm 2.1 while remaining always computationally feasible. The suggestion is as follows. Pick a user specified number N_{max}, say $N_{max} = 50$, and let M be the largest integer for which $\binom{M}{k-1} \leq N_{max}$. In step j of Algorithm 2.1, the critical value is then computed as follows.

$$\hat{d}_{n,A_j}(1-\alpha,k) = \max\{\hat{c}_{n,K}(1-\alpha,k) : K = A_j \cup I,\ I \subset \{r_{\max\{1,|R_j|-M+1\}},\ldots,r_{|R_j|}\},\ |I| = k-1\}\,.$$

That is, we maximize over subsets I not necessarily of the entire index set R_j of previously rejected hypotheses but only of the index set corresponding to the M least significant hypotheses rejected so far. (Of course, when $M \geq |R_j|$, we maximize over all subsets I of R_j of size

$k-1$.) The philosophy of this operative method is to be as close as possible to the generic Algorithm 2.1, given the limitation to the computational burden expressed by N_{max}. The operative method allows for true hypotheses to be among the M least significant hypotheses rejected so far, while the streamlined method assumes they are among the $k-1 < R$ least significant hypotheses rejected so far.

Finally, note that the streamlined algorithm is a special case of the operative method when $N_{max} = 1$ is chosen, and then $M = k-1$. ■

3.2 A General Subsampling Construction

In this subsection, we present an alternative construction of critical values in our stepdown procedure by using subsampling. Unlike the previous subsection, we do not assume H_i is concerned with the test of a parameter θ_i; the approach here is quite general and will hold under weaker asymptotic conditions as well. For any $K \subset \{1, \ldots, s\}$, let $G_{n,K}(P)$ be the joint distribution of the statistics $T_{n,i}$, $i \in K$ under P, with corresponding joint c.d.f. $G_{n,K}(x, P)$, $x \in \mathbb{R}^{|K|}$. Also, let $H_{n,K}(k, P)$ denote the distribution of k-max$(T_{n,i} : i \in K)$ under P. As in Subsection 2.1, let $c_{n,K}(1-\alpha, k, P)$ denote a $1-\alpha$ quantile of $H_{n,K}(k, P)$.

We will make the following general assumption.

Assumption S Under P, the joint distribution of the test statistics $T_{n,i}$, $i \in I(P)$, has a limiting distribution; that is,

$$G_{n,I(P)}(P) \xrightarrow{L} G_{I(P)}(P) \ . \tag{32}$$

This implies that, under P, k-max$(T_{n,i} : i \in I(P))$ has a limiting distribution, say $H_{I(P)}(k, P)$, with limiting c.d.f. $H_{I(P)}(x, k, P)$. Let $c_K(\alpha, k, P)$ denote an α-quantile of $H_{I(P)}(k, P)$; more concretely

$$c_K(\alpha, k, P) = \inf\{x : \ P\{k\text{-max}_{i \in K}(T_{n,i}) \leq x\} \geq \alpha\} \ .$$

We will assume further that

$$H_{I(P)}(x, k, P) \ \text{ is continuous and strictly increasing at } x = c_{I(P)}(1-\alpha, k, P) \ . \tag{33}$$

Note that the continuity condition in (33) is satisfied if the $|I(P)|$ univariate marginal distributions of $G_{I(P)}(P)$ are continuous; see Lemma B.1. Also, the strictly increasing assumption can be removed; see Remark 1.2.1 of Politis et al. (1999). However, it holds in all known examples where the continuity assumption holds, as typical limit distributions are of the Gaussian, Chi-squared, etc. type.

We now detail the general subsampling construction. To this end, assume that we have available an i.i.d. sample $X_1, \ldots, X_n$ from P, and $T_{n,i} = T_{n,i}(X_1, \ldots, X_n)$ is the test statistic we wish to use for testing H_i. To describe the test construction, fix a positive integer $b < n$ let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n := \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Let $T_{b,i}^{(a)}$ be equal to the statistic $T_{b,i}$ evaluated at the data set Y_a, for $a = 1, \ldots, N_n$. Then, for any subset $K \subset \{1, \ldots, s\}$, the joint distribution of $(T_{n,i} : i \in K)$ can be approximated by the empirical distribution of the N_n values $\{T_{b,i}^{(a)} : \ i \in K\}$. In other words, for $x \in \mathbb{R}^s$, the true

joint c.d.f. of the test statistics evaluated at x,

$$G_{n,\{1,\ldots,s\}}(x,P) = P\{T_{n,1} \le x_1, \ldots, T_{n,s} \le x_s\}$$

is estimated by the subsampling distribution

$$\hat{G}_{n,\{1,\ldots,s\}}(x) = \frac{1}{N_n}\sum_a I\{T_{b,1}^{(a)} \le x_1, \ldots, T_{b,s}^{(a)} \le x_s\} \ . \tag{34}$$

Note that the marginal distribution of any subset $K \subset \{1,\ldots,s\}$, $G_{n,K}(P)$, is then approximated by the marginal distribution induced by (34) on that subset of variables. So, $\hat{G}_{n,K}$ refers to the empirical distribution of the values $\{T_{n,i}^{(a)} : \ i \in K\}$. (In essence, one only has to estimate one joint sampling distribution for all the test statistics because this then induces that of any subset, even though we are not assuming anything like subset pivotality.)

Similarly, the estimate of the whole joint distribution of test statistics induces an estimate for the distribution of the maximum or kth largest of test statistics. Specifically, $H_{n,K}(k,P)$ is estimated by the empirical distribution $\hat{H}_{n,K}(x,k)$ of the values k-max$(T_{n,i}^{(a)} : \ i \in K)$; that is,

$$\hat{H}_{n,K}(x,k) = \frac{1}{N_n}\sum_a I\{k\text{-max}(T_{b,i}^{(a)} : \ i \in K) \le x\} \ .$$

Also, let

$$\hat{c}_{n,K}(1-\alpha,k) = \inf\{x : \hat{H}_{n,K}(x,k) \ge 1-\alpha\} \tag{35}$$

denote the estimated $1-\alpha$ quantile of the k-max of test statistics $T_{n,i}$ with $i \in K$.

Note the monotonicity of the critical values: for $I \subset K$

$$\hat{c}_{n,K}(1-\alpha,k) \ge \hat{c}_{n,I}(1-\alpha,k) \ . \tag{36}$$

This simple observation together with Theorem 2.1 immediately implies the following.

Corollary 3.2 *Under the setup and notation of this subsection, consider Algorithm 2.1 with critical values given by (35). Then,*

$$k\text{-}FWER_P \le P\{k\text{-max}(T_{n,i} : \ i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha,k)\} \ . \tag{37}$$

The following result proves consistency and strong control of our stepdown algorithm based on these subsample estimates of critical values. Note, in particular, that Assumption B2 is not needed here at all, a reflection of the fact that the bootstrap requires much stronger (local uniform convergence) assumptions for consistency; see Politis et al. (1999). Also notice that we do not even need to assume that there exists a P for which all hypotheses are true.

Theorem 3.4 *Suppose Assumption S holds. Let $b/n \to 0$ and $b \to \infty$.*

(i) The subsampling approximation satisfies

$$\rho\left(\hat{G}_{n,I(P)}, G_{n,I(P)}(P)\right) \xrightarrow{P} 0 \ , \tag{38}$$

for any metric ρ metrizing weak convergence on $\mathbb{R}^{|I(P)|}$.

(ii) The subsampling critical values satisfy

$$\hat{c}_{n,I(P)}(1-\alpha,k) \xrightarrow{P} c_{I(P)}(1-\alpha,k) \ . \tag{39}$$

(iii) Therefore, using Algorithm 2.1 with $\hat{c}_{n,K}(1-\alpha,k)$ given by (35) results in $\limsup_n k\text{-}FWER_P \leq \alpha$.

Example 3.3 (Cube root asymptotics) Kim and Pollard (1990) show that a general class of M-estimators converge at rate $\tau_n = n^{1/3}$ to a non-normal limiting distribution. As a result, inconsistency of the bootstrap typically follows; see Abrevaya and Huang (2005). On the other hand, Delgado et al. (2001) demonstrate the consistency of the subsampling method for constructing hypothesis tests for a single null hypothesis. By similar arguments, the validity of the subsampling construction of Theorem 3.4 in the context of cube root asymptotics can be established. ■

The above approach can be extended to dependent data. For example, if the data $X_1, \ldots, X_n$ form a stationary time series, we would only consider the $n-b+1$ subsamples of the form $(X_a, X_{a+1}, \ldots, X_{a+b-1})$. Generalizations for nonstationary time series, random fields, and point processes are further treated in Politis et al. (1999).

4 Asymptotic Results on FDP Control

In some applications, one might be willing to tolerate a larger number of false rejections in case the total number of rejections is large. In other words, one might be willing to tolerate a certain (small) fraction of false rejections out of the total rejections. This leads to control based on the *false discovery proportion* (FDP). Let F be the number of false rejections made by a multiple testing procedure and let R be the total number of rejections. Then the FDP is defined as follows:

$$\text{FDP} = \begin{cases} \frac{F}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

Control of the FDP corresponds to the control of $P\{\text{FDP} > \gamma\}$ where γ is a user specified number $\gamma \in [0,1)$. A typical value may be $\gamma = 0.1$; the choice $\gamma = 0$ corresponds to control of the FWER.

A multiple testing procedure is said to control the FDP at level α if, for the given sample size n, $P\{\text{FDP} > \gamma\} \leq \alpha$, for all P. A multiple testing procedure is said to asymptotically control the FDP at level α, if $\limsup_n P\{\text{FDP} > \gamma\} \leq \alpha$, for all P. Our focus will be on procedures that provide asymptotic control.

Notice that a procedure satisfying $P\{\text{FDP} > \gamma\} \leq 0.5\}$ guarantees that the median of the FDP is $\leq \gamma$. Of course, this is distinct from control of the mean of the FDP, which is the FDR, though they are similar in spirit.

The approach we propose is built upon an underlying procedure that (asymptotically) controls the k-FWER for any fixed $k \geq 1$. We then sequentially apply this k-FWER procedure

for $k = 1, 2, \ldots$ until a stopping rule indicates termination. In the end, we reject all hypotheses that were rejected in the last round of applying the k-FWER procedure.

To develop the idea, consider controlling $P\{\text{FDP} > 0.1\}$. We start out by applying the 1-FWER procedure, that is, by (asymptotically) controlling the FWER. Denote by N_1 the number of hypotheses rejected. Due to the FWER control, one can be confident that no false rejection has occurred and that, in return, the FDP has been controlled. Consider now rejecting $H_{r_{N_1+1}}$, the next most significant hypothesis. Of course, if $H_{r_{N_1+1}}$ is false, there is nothing to worry about, so suppose $H_{r_{N_1+1}}$ is true. In case the FWER was controlled successfully in the first step, the FDP upon rejection of $H_{r_{N_1+1}}$ then becomes $1/(N_1 + 1)$, which is greater than 0.1 if and only if $N_1 < 9$. So if $N_1 \geq 9$ we can reject one true hypothesis and still avoid FDP > 0.1. This suggests to stop if $N_1 < 9$ and otherwise to apply the 2-FWER procedure which, by design, controls the probability of making two or more false rejections. Denote the total number of hypotheses rejected by the 2-FWER base procedure by N_2. Reasoning similarly to before, if $N_2 < 19$, we stop and otherwise we apply the 3-FWER procedure. If N_j denotes the total number of hypotheses rejected by the j-FWER procedure, the stepdown method is continued until $N_j < 10j - 1$, at which point termination incurs.

The following algorithm summarizes the method for arbitrary γ.

Algorithm 4.1 (Generic Method for Control of the FDP)

1. Let $j = 1$ and let $k_1 = 1$.
2. Apply the k_j-FWER procedure and denote by N_j the number of hypotheses it rejects.
3. (a) If $N_j < k_j/\gamma - 1$, stop and reject all hypotheses rejected by the k_j-FWER procedure.
   (b) Otherwise, let $j = j + 1$ and then $k_j = k_{j-1} + 1$. Return to step 2.

Note that the algorithm does not assume anything on the nature of the underlying k-FWER procedure; for example, the procedures of Lehmann and Romano (2005a) or a single-step procedure along the lines of Subsection 2.1 could be used. However, in order to reject as many false hypotheses as possible while maintaining (asymptotic) control of the FDP, we suggest to employ a stepwise procedure that accounts for the dependence structure of the test statistics $T_{n,i}$.

Algorithm 4.1 is similar to the proposal of Korn et al. (2004) for FDP control which is, however, restricted to a multivariate permutation model. The proposal of Korn et al. (2004) is heuristic in the sense that they cannot guarantee finite sample nor asymptotic control of the FDP even if the permutation hypothesis is valid. In our study of Algorithm 4.1, even if the k-FWER base procedure provides finite sample control for any k, we cannot guarantee finite sample control of the FDP. However, we will provide arguments for asymptotic control. (Also, our simulations presented later show good finite sample control.) The theorem below considers a general bootstrap construction where the individual tests are one-sided and concern univariate parameters $\theta_i(P)$. The bootstrap construction for two-sided tests and the more general subsampling construction can be handled similarly; the details are left to the reader.

Theorem 4.1 *Consider the setup of Theorem 3.1. Fix P satisfying Assumption B3. Employ the stepdown procedure of Algorithm 2.1 with $\hat{c}_{n,K}(1-\alpha,k)$ replaced by $b_{n,K}(1-\alpha,\hat{Q}_n,k)$ as the underlying k-FWER procedure. Then the following statements concerning Algorithm 4.1 are true.*

(i) If P is such that $i \notin I(P)$, i.e. H_i is false and $\theta_i(P) > 0$, then the probability that the method rejects H_i tends to one.

(ii) $\limsup_n P\{FDP > \gamma\} \leq \alpha$.

Remark 4.1 The theorem remains valid if the bootstrap k-FWER procedure is based on the operative method of Remark 3.7 or the streamlined Algorithm 2.2 instead of the generic Algorithm 2.1. But, again, in view of finite sample performance, we suggest the use of the generic Algorithm 2.1 if feasible or at least the use of the operative method. ■

5 Comparison With Related Methods

We have proposed stepdown procedures that control the k-FWER and the FDP, with the goal of improving upon methods that do not attempt to incorporate or estimate the dependence structure between the test statistics or p-values. An alternative approach toward achieving this goal is given in van der Laan et al. (2004). We briefly discuss their proposal.

The approach of van der Laan et al. (2004) begins with a procedure that controls the 1-FWER (i.e., the usual FWER) and then rejects in addition the $k-1$ most significant hypotheses not rejected so far. They coin this an *augmentation procedure*, since the 1-FWER rejection set is augmented by the $k-1$ next most significant hypotheses to arrive at the k-FWER rejection set. Obviously, if the 1-FWER procedure succeeds in (asymptotically) controlling the 1-FWER, then the augmented procedure provides (asymptotic) control of the k-FWER. However, this approach seems suboptimal, because it makes the worst case assumption that, having achieved 1-FWER control, the $k-1$ next most significant hypotheses are all true hypotheses. Moreover, $k-1$ additional hypotheses are always rejected, even if the test statistics or p-values to which they correspond are clearly not significant. (In fact, one can reject any $k-1$ additional hypotheses, not just the next $k-1$ most significant ones.) In addition, the approach really does not fully utilize the weaker measure of error control afforded by using the k-FWER with $k > 1$, in that the augmentation method will reject more than $k-1$ hypotheses if and only if the 1-FWER controlling procedure rejects some hypotheses, and this criterion may be too strong to admit any rejections.

Our approach to control the k-FWER is based on knowing or estimating the sampling distribution of a suitable k-max statistic, that is, the kth largest of the s individual (possibly standardized) test statistics. A hypothesis is rejected if its corresponding test statistic is large (relative to the the estimated quantiles of the sampling distribution of the k-max statistic), unlike the augmentation approach where a hypothesis can be rejected even if its corresponding test statistic is not deemed large by any measure.

To appreciate how the two approaches differ, first consider augmentation based on the Holm procedure, given by (7) with $k = 1$. Other than the additional $k - 1$ hypotheses that are rejected after applying Holm, the procedure can only reject a nontrivial number (k or more) if and only if the smallest p-value is $\leq \alpha/s$. On the other hand, the generalized Holm procedure starts out with a great advantage; the smallest p-value is compared with $k\alpha/s$, a k-fold increase! While it is possible for augmentation to reject more hypotheses, it can only reject $k-1$ more than the generalized Holm procedure (and these additional rejections may be suspect because they can correspond to large p-values), but the generalized Holm procedure can reject many, many more.

Similar comparisons can be made with augmentation applied to a FWER controlling procedure that attempts to account for the dependence structure (like the ones in this paper with $k = 1$). Augmentation can possibly reject $k-1$ more hypotheses than the ones we propose here, but our methods can easily reject many more. Note that, if the test statistics or p-values are independent, then augmentation of a bootstrap method that controls the FWER still cannot produce anything much better than the Holm method; similarly, the resampling method here cannot procedure anything much better than the generalized Holm method. (To appreciate why, see Problem 9.2 of Lehmann and Romano (2005b).) Thus, in the case of independence, the two methods essentially behave as described in the previous paragraph.

The comparison is similar for the procedures (asymptotically) controlling the FDP. Our approach is to sequentially apply a k-FWER procedure for $k = 1, 2, \ldots$ until a stopping rule indicates termination. On the other hand, van der Laan et al. (2004) again augment the rejection set of a 1-FWER procedure. The idea now is as follows. Let R denote the number of rejections by the 1-FWER procedure. Then reject in addition the D next most significant hypotheses where D is the largest integer which satisfies

$$\frac{D}{D+R} \leq \gamma$$

Again, if the 1-FWER procedure succeeds in (asymptotically) controlling the 1-FWER, then the augmented procedure provides (asymptotic) control of the FDP. But also again, this approach seems pessimistic in that it makes the worst case assumption that, having achieved 1-FWER control, the D next most significant hypotheses are all true hypotheses.

The next section compares the finite sample performance of the two approaches.

6 Simulation Study

This section presents a small simulation study in the context of testing population means. We generate random vectors $X_1, \ldots, X_n$ from an s-dimensional multivariate normal distribution with mean vector $\theta = (\theta_1, \ldots, \theta_s)$, where $n = 100$ and $s = 50$ or $s = 400$. The null hypotheses are $H_i : \theta_i \leq 0$ and the alternative hypotheses are $H_i : \theta_i > 0$. The test statistics are $T_{n,i} = \sqrt{n}\bar{X}_{i,\cdot}/S_i$, where

$$\bar{X}_{i,\cdot} = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}, \qquad S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{i,j} - \bar{X}_{i,\cdot})^2 \ ,$$

We consider three models for the covariance matrix Σ having (i,j) component $\sigma_{i,j}$. The models share the feature $\sigma_{i,i} = 1$ for all i; so we are left to specify $\sigma_{i,j}$ for $i \neq j$.

- Common correlation: $\sigma_{i,j} = \rho$, where $\rho = 0$ or $\rho = 0.5$.
- Power structure: $\sigma_{i,j} = \rho^{|i-j|}$ where $\rho = 0.9$.
- Two class structure: the variables are grouped in two classes of equal size $s/2$. Within each class, there is a common correlation of $\rho = 0.5$; and across classes, there is a common correlation of $\rho = -0.5$. Formulated mathematically, for $i \neq j$,

$$\sigma_{i,j} = \begin{cases} 0.5 & \text{if both } i,j \in \{1,\ldots,s/2\} \text{ or both } i,j \in \{s/2+1,\ldots,s\} \\ -0.5 & \text{otherwise} \end{cases},$$

We consider six scenarios for the mean vector $\theta = (\theta_1, \ldots, \theta_s)$. When specifications appear in parentheses, they are for $s = 400$ while the specifications before the parentheses are for $s = 50$.

- All $\theta_i = 0$
- Ten (one hundred) of the $\theta_i = 0.25$ and the remaining $\theta_i = 0$ in one of the following two ways:
  - Every fifth (fourth) $\theta_i = 0.25$; called equal-spaced scenario.
  - There are two (four) blocks $(0.5, 0.5, \ldots, 0.5)$ of size five (twenty-five) within θ; called block scenario.
- Twenty-five (two hundred) of the $\theta_i = 0.25$ and the remaining $\theta_i = 0$ in one of the following two ways:
  - Every other $\theta_i = 0.25$; called equal-spaced scenario.
  - There are five (eight) blocks $(0.5, 0.5, \ldots, 0.5)$ of size five (twenty-five) within θ; called block scenario.
- All $\theta_i = 0.25$

When the covariance matrix is of the common correlation form, then we do not have to distinguish between the equal-spaced and the block scenarios for the mean vector, so in this case there are a total of four scenarios only.

We include the following multiple testing procedures in the study. The value of k is $k = 3$ when $s = 50$ and $k = 10$ when $s = 400$. The nominal level is $\alpha = 0.05$, unless indicated otherwise.

- The bootstrap 1-FWER construction of Subsection 3.1. This procedure is denoted by 1-Boot.
- The k-FWER augmentation procedure of van der Laan et al. (2004), based on the 1-Boot construction. This procedure is denoted by k-Aug.

- The k-FWER generalized Holm procedure described by (7), where the individual p-values are derived from $T_{n,i} \sim t_{n-1}$ under $\theta_i = 0$. This procedure is denoted by k-gH.

- The bootstrap k-FWER construction of Subsection 3.1. This procedure is based on the operative method with $N_{max} = 50$, see Remark 3.7, and is denoted by k-Boot.

- The FDP augmentation procedure of van der Laan et al. (2004) with $\gamma = 0.1$, based on the 1-Boot construction. This procedure is denoted $\text{Aug}_{0.1}$.

- The FDP procedure of Lehmann and Romano (2005a) with $\gamma = 0.1$; see (8). This procedure is denoted $\text{LR}_{0.1}$.

- The bootstrap FDP construction of Section 4 with $\gamma = 0.1$. This procedure is denoted by $\text{Boot}_{0.1}$.

- The bootstrap FDP construction of Section 4 with $\gamma = 0.1$ but nominal level $\alpha = 0.5$. Therefore, this procedure (asymptotically) controls the median FDP to be bounded above by $\gamma = 0.1$; it is denoted by $\text{Boot}_{0.1}^{Med}$.

The performance criteria are (1) the empirical k-FWERs and FDPs, compared to the nominal level $\alpha = 0.05$ (or $\alpha = 0.5$ for the method controlling the median FDP); and (2) the average number of false hypotheses rejected. Since the k-Aug procedure rejects the $k-1$ most significant hypotheses regardless of the data, we also follow this route for the k-gH and k-Boot procedures to ensure a fair comparison as far as (2) is concerned. (Though the differences are really negligible if this route is not followed for the k-gH and k-Boot procedures.) The results are presented in Tables 1–3 for $s = 50$ and in Tables 4–6 for $s = 400$. They can be summarized as follows.

- All methods provide satisfactory finite sample control of their respective k-FWER or FDP criteria. In particular, the finite sample control does not appear to deteriorate when the number of hypotheses is increased from $s = 50$ to $s = 400$, while the sample size is kept fixed at $n = 100$.

- Depending on context, our stepwise k-FWER procedure can detect many more false alternatives compared to the 1-FWER procedure. The same is not true for the augmentation procedure of van der Laan et al. (2004), since, by design, it detects at most $k-1$ more false hypotheses compared to the 1-FWER procedure. So especially when s is large, this approach appears suboptimal. Even the conservative k-FWER generalized Holm method k-gH, which is based on individual p-values and does not even attempt to account for the dependence across the p-values, is more powerful than the augmentation method for large s.

- The comparison is similar for the various FDP procedures with $\alpha = 0.05$. Our bootstrap procedure is the most powerful one. The augmentation procedure of van der Laan et al. (2004) becomes uncompetitive when s is large. Even the conservative FDP method of

Lehmann and Romano (2005a), which is based on individual p-values and does not even attempt to account for the dependence across the p-values, is often more powerful than the augmentation procedure for large s.

The procedure controlling the median FDP (last column) is always the most powerful one. However, it should be understood that it is philosophically different from the other FDP procedures. If $P\{\text{FDP} > 0.1\} \leq 0.05$ is achieved, then, in a given application, one can be 95% confident that the realized FDP is at most 0.1. On the other hand, if $P\{\text{FDP} > 0.1\} \leq 0.5$ is achieved (i.e., control of the median FDP), then, in a given application, one can only by 50% confident that the realized FDP is at most 0.1. So, loosely speaking, there is a good chance that the realized FDP ends up greater than 0.1, and perhaps by quite a bit.

To examine this issue, we look at the sampling distribution of the FDP when the median FDP is controlled. Four scenarios are considered for $s = 50$, namely scenarios 2, 3, 6, and 7 of Table 1; and four scenarios are considered for $s = 400$, namely scenarios 2, 3, 6, and 7 of Table 4. Figure 1 summarizes the distribution of the corresponding realized FDPs via boxplots. It can be seen that, while median FDP control is achieved, the variation of the sampling distributions is considerable, in particular for the case of common correlation $\sigma_{i,j} = 0.5$. As a result, the realized FDP may well be quite above $\gamma = 0.1$.

A similar problem arises in controlling the false discovery rate (FDR), as proposed by Benjamini and Hochberg (1995). The FDR is the expected value of the FDP. Like the median FDP, it is also a measure of central tendency of the sampling distribution of the FDP. In a given application, the realized FDP can be quite far away from its expected value, the FDR, as made clear in Korn et al. (2004).

7 Concluding Remarks.

We have shown how computationally feasible stepdown methods can be constructed to control generalized error rates in multiple testing. On the one hand, we have considered the k-FWER, which is defined as the probability of making k or more false rejections. This concept would be appropriate when a given number of false rejections can be tolerated. On the other hand, we have also considered the FDP, which is the ratio of false rejections out of the total number of rejections (and defined to be zero when there are no rejections). This concept would be appropriate when a certain proportion of false rejections can be tolerated. Some simulations have shown that these less strict methods can reject many more false hypotheses compared to the traditional FWER control, especially when the number of hypotheses under test is large.

Our stepdown methods (asymptotically) account for the dependence structure across test statistics. As a result, they are more powerful than the generalized Holm stepdown methods of Hommel and Hoffman (1987) and Lehmann and Romano (2005a), which are based on individual p-values and designed to handle a 'worst case' dependence structure. An alternative approach that also accounts for the dependence structure across test statistics is the augmentation approach of van der Laan et al. (2004). However, simulations show their methods are noticeable less powerful, especially when the number of hypotheses under test is large.

A A Generalized Closure Method.

In this section, we generalize the closure method of Marcus et al. (1976) to obtain methods that control the k-FWER. As before, we are testing s hypotheses $H_1, \ldots, H_s$ based on X from some probability distribution $P \in \Omega$; H_i asserts $P \in \omega_i$. For fixed k, $K \subset \{1, \ldots, s\}$ with $|K| \geq k$, let $H_{K,k}$ be the hypothesis that all H_i, $i \in K$, are true, except possibly at most $k-1$ of them. More formally, if $d = |K| - (k-1)$, then

$$H_{K,k} = \bigcup \left\{ \{ \bigcap_{j=1}^{d} \omega_{i_j} \} : \ i_1, \ldots, i_d \text{ distinct indices in } K \right\} .$$

Suppose an α level test is available to test $H_{K,k}$, i.e. a single test that controls the usual probability of a Type 1 error. The generalized closed testing method rejects any hypothesis H_i if and only if $H_{K,k}$ is rejected whenever $i \in K$ and $|K| \geq k$.

Theorem A.1 *The above testing method controls the k-FWER at level α if the probability of a Type 1 error is $\leq \alpha$ when testing $H_{K,k}$.*

Proof. Let $I = I(P)$ be the set of indices of true hypotheses. Assume $|I| \geq k$ or there is nothing to prove. Define the events

$$A = \{\text{at least } k \text{ true hypotheses rejected}\}$$

and

$$B = \{H_{I,k} \text{ rejected}\} .$$

By the description of the closed testing method, the event A implies B, and so $A = A \cap B$. Therefore,

$$k\text{-FWER}_P = P\{A\} = P\{A \cap B\} = P\{B\}P\{A|B\} \leq P\{B\} \leq \alpha . \blacksquare$$

The value of the generalized closed testing method is that the problem of controlling the k-FWER is reduced to the problem of controlling the usual probability of a Type 1 error of single tests. However, in order to carry out the procedure, one must essentially carry out tests of $H_{K,k}$ for all K with $|K| \geq k$. Typically, k is small compared with s, and so the number of such tests can be nearly 2^s. The main point of the present work is to show that one can carry out this closure method in a computationally feasible manner if we have monotonicity of critical values. In the body of our paper, all tests were based on the k-max statistic. Theorem 2.1 can apply more generally.

B Proofs

Proof of Theorem 2.1 Assume any configuration of true and false null hypotheses with $|I(P)| \geq k$, or there is nothing to prove. Consider the event that at least k true null hypotheses

are rejected, so that for at least k indices $i \in I(P)$, hypothesis H_i is rejected. Let $\hat{j}$ be the (random) smallest index j in the algorithm where this occurs, so that

$$k\text{-max}(T_{n,i} : \ i \in I(P)) > \hat{d}_{n,A_{\hat{j}}}(1-\alpha, k) \ , \tag{40}$$

By definition of $\hat{j}$ (now fixed),

$$I(P) \subset A_{\hat{j}} \cup I_0 \ ,$$

where I_0 is some set of indices satisfying $I_0 \subset R_{\hat{j}}$ and $|I_0| = k-1$. Let L be any set of indices of false null hypotheses (not necessarily uniquely defined) which satisfy

$$A_{\hat{j}} \cup I_0 = I(P) \cup L \ .$$

Since $\hat{d}_{n,A_{\hat{j}}}(1-\alpha, k)$ is defined by taking the maximum over sets I of $\hat{c}_{n,K}(1-\alpha, k)$ with $K = A_{\hat{j}} \cup I$ as I varies over indices satisfying $I \subset R_{\hat{j}}$ and $|I| = k-1$, it follows that

$$\hat{d}_{n,A_{\hat{j}}}(1-\alpha, k) \geq \hat{c}_{n,I(P)\cup L}(1-\alpha, k) \ .$$

By the monotonicity assumption,

$$\hat{c}_{n,I(P)\cup L}(1-\alpha, k) \geq \hat{c}_{n,I(P)}(1-\alpha, k) \ .$$

To summarize, the event that at least k true null hypotheses are rejected implies that

$$k\text{-max}(T_{n,i} : i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha, k)$$

and so (i) follows. Part (ii) follows immediately from (i). ■

Lemma B.1 *Let $k \leq s$.*

(i) The k-max function is continuous; that is, if $y_n = (y_{n,1}, \ldots, y_{n,s}) \in \mathbb{R}^s$ and $y_n \to y \in \mathbb{R}^s$, then, as $n \to \infty$,

$$k\text{-}\max(y_{n,1}, \ldots, y_{n,s}) \to k\text{-}\max(y_1, \ldots, y_s) \ .$$

(ii) If $Y_n \in \mathbb{R}^s$ and $Y_n \xrightarrow{L} Y$, then

$$k\text{-}\max(Y_{n,1}, \ldots, Y_{n,s}) \xrightarrow{L} k\text{-}\max(Y_1, \ldots, Y_s) \ .$$

(iii) Furthermore, if each Y_i in (ii) has a continuous marginal distribution, then the distribution of k-$\max(Y_1, \ldots, Y_s)$ is continuous.

Proof of Lemma B.1 Part (i) is trivial, and the continuous mapping theorem then implies (ii). To prove (iii), for any $x \in \mathbb{R}$,

$$P\{k\text{-max}(Y_1, \ldots, Y_s) = x\} \leq \sum_{i=1}^{s} P\{Y_i = x\} \ . \ \blacksquare$$

Proof of Theorem 3.1

To prove (i), note that by Corollary 3.1 it is sufficient to show that

$$\limsup_n P\{k\text{-max}(T_{n,i}:\ i \in I(P)) > b_{n,I(P)}(1-\alpha, k, \hat{Q}_n)\} \leq \alpha \tag{41}$$

Since $\theta_i(P) \leq 0$ for $i \in I(P)$, it follows that

$$k\text{-max}(T_{n,i}:\ i \in I(P)) = k\text{-max}(\tau_n\hat{\theta}_{n,i}:\ i \in I(P)) \leq k\text{-max}(\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]:\ i \in I(P))\ .$$

Therefore, the left side of (41) is bounded above by

$$\lim_n P\{k\text{-max}(\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]:\ i \in I(P)) > \hat{b}_{n,I(P)}(1-\alpha, k, \hat{Q}_n)\}\ . \tag{42}$$

Assumptions B1 and B2 together with the continuous mapping theorem imply that

$$\rho\left(L_{n,I(P)}(k,P), L_{n,I(P)}(k,\hat{Q}_n)\right) \xrightarrow{P} 0\ ,$$

for any metric ρ metrizing weak convergence on $\mathbb{R}$. Hence, it follows that (42) is equal to α, by an argument very similar to the proof of Theorem 1 of Beran (1984).

To prove (ii), assume $\theta_i(P) > 0$. Since Assumptions B1 and B2 continue to hold when K is replaced by $A_1 = \{1, \ldots, s\}$, $b_{n,A_1}(1-\alpha, k, \hat{Q}_n)$ is stochastically bounded. Furthermore, by the continuous mapping theorem, $\tau_n[\hat{\theta}_{i,n} - \theta_i(P)]$ has a limiting distribution, so $T_{n,i} = \tau_n\hat{\theta}_{i,n} \xrightarrow{P} \infty$. Therefore, with probability tending to one, $T_{n,i} > b_{n,A_1}(1-\alpha, k, \hat{Q}_n)$, resulting in the rejection of H_i in the first step of Algorithm 2.1. ■

Proof of Theorem 3.2 The proof is completely analogous to the proof of Theorem 3.1. The only additional fact needed to prove (iii) is that, when $\theta_i(P) > 0$, $\tau_n\hat{\theta}_{n,i} > 0$ with probability tending to one, and similarly for $\theta_i(P) < 0$. Indeed, assumption B1(i) implies $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ has a limiting distribution, which implies $\tau_n\hat{\theta}_{n,i} \xrightarrow{P} \infty$ when $\theta_i(P) > 0$, and $\tau_n\hat{\theta}_{n,i} \xrightarrow{P} -\infty$ when $\theta_i(P) < 0$. ■

Proof of Theorem 3.3

To prove (i), assume $\theta_i(P) > 0$. Since Assumptions B1 and B2 continue to hold when K is replaced by $A_1 = \{1, \ldots, s\}$, $b_{n,A_1}(1-\alpha, k, \hat{Q}_n)$ is stochastically bounded. Furthermore, by the continuous mapping theorem, $\tau_n[\hat{\theta}_{i,n} - \theta_i(P)]$ has a limiting distribution, so $T_{n,i} = \tau_n\hat{\theta}_{i,n} \xrightarrow{P} \infty$. Therefore, with probability tending to one, $T_{n,i} > b_{n,A_1}(1-\alpha, k, \hat{Q}_n)$, resulting in the rejection of H_i in the first step of Algorithm 2.2.

To prove (ii), note that by reasoning similar to before, $\min(T_{n,i} : i \notin I(P)) \xrightarrow{P} \infty$. On the other hand, $\max(T_{n,i} : i \in I(P))$ is either bounded in probability, in case $\theta_i(P) = 0$ for at least one $i \in I(P)$, or $\max(T_{n,i} : i \in I(P)) \xrightarrow{P} -\infty$, in case $\theta_i(P) < 0$ for all $i \in I(P)$.Therefore, the event

$$\min(T_{n,i} : i \notin I(P)) > \max(T_{n,i} : i \in I(P)) \tag{43}$$

has probability tending to one. But if the event (43) happens, then the rejected true hypotheses (if such exist) will always be the least significant hypotheses among the rejected hypotheses at

any stage. This together with the monotonicity of the critical values $b_{n,K}(1-\alpha, k, \hat{Q}_n)$ allows us to follow asymptotic control of the k-FWER from (41) even when Algorithm 2.2 is used. But (41) was already established in the proof of Theorem 3.1. ■

Proof of Theorem 3.4

The proof of (i) is the essential subsampling argument, which derives from (34) being a U-statistic; see Politis et al. (1999), Theorem 2.6.1, where one statistic is treated, but the argument is extendable to the simultaneous estimation of the joint distribution. The result (ii) follows as well.

To prove (iii), note that by Corollary 3.2 it is sufficient to show that

$$\limsup_n P\{k\text{-max}(T_{n,i}: \ i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha, k)\} \leq \alpha \tag{44}$$

But part (ii) of the Theorem implies, for any $\epsilon > 0$,

$$\hat{c}_{n,I(P)}(1-\alpha, k) \geq c_{I(P)}(1-\alpha, k) - \epsilon \quad \text{with probability } \to 1 \ .$$

Therefore, using Assumption S, the limit superior of the probability of violation of the k-FWER criterion is bounded above as follows, for any $\epsilon > 0$,

$$\limsup_n k\text{-FWER}_P \leq P\{k\text{-max}(T_i, \ i \in I(P)) > c_{I(P)}(1-\alpha) - \epsilon\} \ ,$$

where $(T_i, \ i \in I(P))$ denote variables whose joint distribution is $G_{I(P)}(P)$. But letting $\epsilon \to 0$, the right side of the last expression becomes

$$1 - H_{I(P)}(c_{I(P)}(1-\alpha), P) = 1 - (1-\alpha) = \alpha \ .$$

Proof of Theorem 4.1

The proof of (i) follows immediately from part (ii) of Theorem 3.1.

To prove (ii), note that by reasoning similar to the proof of part (ii) of Theorem 3.3, with probability tending to one, all false hypothesis are rejected before any true hypothesis comes under scrutiny. Therefore, with probability tending to one, a violation of the FDP criterion occurs if and only if the event

$$F > \frac{\gamma}{1-\gamma}(s - |I(P)|) \tag{45}$$

occurs, where F is the number of true hypotheses rejected by Algorithm 4.1. Let $F(k)$ denote the number of true hypotheses rejected by the bootstrap k-FWER procedure. Furthermore, let k^* denote the smallest integer greater than $(\gamma/(1-\gamma)(s - |I(P)|)$. Assume $|I(P)| \geq k^*$ or there is nothing to prove. By the above argument, we therefore have

$$\begin{aligned} \limsup_n P\{\text{FDP} > \gamma\} &= \limsup_n P\{F \geq k^*\} \\ &\leq \limsup_n P\{F(k^*) \geq k^*\} \\ &\leq \alpha \quad \text{(by part (ii) of Theorem 3.1).} \end{aligned} \tag{46}$$

To see that (46) holds true, note the following two facts. First, the bootstrap k-FWER procedure is monotone in k: any hypothesis rejected by the k_1-FWER procedure will also be rejected by the k_2-FWER procedure as long as $k_1 < k_2$. Second, according to step 3.(a) of Algorithm 4.1, the algorithm terminates with the application of the k^*-FWER procedure, or even before then, if

$$N_{k^*} < \frac{k^*}{\gamma} - 1 \tag{47}$$

In case all false hypotheses are rejected first, the event (47) happens if and only if

$$k^* > \frac{\gamma}{1-\gamma}(s - I(P) - [F(k^*) - (k^* - 1)]) \ . \tag{48}$$

By the definition of k^*, the inequality (48) will hold as long as $F(k^*) \leq k^* - 1$. Therefore, the event $F(k^*) \leq k^* - 1$ implies that (1) $F(k) \leq k^* - 1$ for any $k < k^*$; and that (2) Algorithm 4.1 terminates with the application of the k^*-FWER procedure, or even before then, if all false hypotheses are rejected first (which happens with probability tending to one). These two facts together demonstrate the validity of (46). ■

References

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*. Forthcoming.

Aitken, M. (1969). Some tests for correlation matrices. *Biometrika*, 56:443–446.

Aitken, M. (1971). Correction to 'some tests for correlation matrices'. *Biometrika*, 58:245.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresberichte des Deutschen Mathematischen Vereins*, 86:14–30.

Beran, R. (1986). Simulated power functions. *Annals of Statistics*, 14:151–173.

Beran, R. (1988a). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83:679–686.

Beran, R. (1988b). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697.

Delgado, M., Rodríguez-Poo, J., and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters*, 73:241–250.

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13. Available at `http://www.bepress.com/sagmb/vol3/iss1/art13`.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Annals of Statistics*, 27:274–289.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47:757–762.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23:169–192.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Hommel, G. and Hoffman, T. (1987). Controlled uncertainty. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypthesis Testing*, pages 154–161. Springer, Heidelberg.

Kim, J. and Pollard, D. B. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Lehmann, E. L. and Romano, J. P. (2005a). Generalizations of the familywise error rate. *Annals of Statistics*, 33. Forthcoming.

Lehmann, E. L. and Romano, J. P. (2005b). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Annals of Statistics*, 33. Forthcoming.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660.

Perone Pacifico, M., Genovese, C. R., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014.

Petrondas, D. and Gabriel, K. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association*, 78(384):949–957.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Rogers, J. and Hsu, J. (2001). Multiple comparisons of biodiversity. *Biometrical Journal*, 43:617–625.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Romano, J. P. and Shaikh, A. M. (2004). On control of the false discovery proportion. Technical Report 2004-31, Department of Statistics, Stanford University.

Romano, J. P. and Shaikh, A. M. (2005). Stepup procedures for control of generalizations of the familywise error rate. Technical Report 2005-2, Department of Statistics, Stanford University.

Romano, J. P. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*. Forthcoming.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30(1):239–257.

Troendle, J. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association*, 90:370–378.

Tu, W. and Zhou, X. (2000). Pairwise comparison of the means of skewed data. *Journal of Statistical Planning and Inference*, 88:59–74.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15. Available at `http://www.bepress.com/sagmb/vol3/iss1/art15/`.

Westfall, P. H. and Wolfinger, R. D. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51:3–8.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley, New York.

Westfall, P. H., Zaykin, D. V., and Young, S. S. (2001). Multiple tests for genetic effects in association studies. In Looney, S., editor, *Methods in Molecular Biology: Biostatistical Methods*, volume 184, pages 143–168. Humana Press, Toloway, NJ.

Table 1: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 200$.

Common correlation: $\sigma_{i,j} = 0$								
	1-Boot	3-Aug	3-gH	3-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.4	5.4	0.0	4.5	5.4	4.7	5.4	51.1
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_i = 0.25$								
Control	4.5	0.0	0.0	2.9	4.5	4.1	4.5	49.0
Rejected	2.7	4.5	3.9	6.3	2.7	2.6	2.7	6.4
Twenty-five $\theta_i = 0.25$								
Control	3.2	0.0	0.0	2.0	1.6	1.7	2.6	38.6
Rejected	7.0	9.0	9.5	16.7	7.3	7.2	7.9	21.3
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	15.1	17.1	19.2	41.6	16.5	15.4	44.9	50.0
Common correlation: $\sigma_{i,j} = 0.5$								
	1-Boot	3-Aug	3-gH	3-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.5	5.5	1.7	5.4	5.5	3.0	5.5	50.3
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_i = 0.25$								
Control	5.0	3.0	1.3	4.7	4.2	2.3	4.9	49.2
Rejected	3.4	5.3	4.3	5.7	3.5	2.7	3.5	8.3
Twenty-five $\theta_i = 0.25$								
Control	5.4	2.4	0.1	4.6	3.2	1.6	4.6	48.9
Rejected	9.2	11.1	9.9	14.4	9.7	8.2	10.8	22.8
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	10.5	22.5	19.2	32.8	21.8	22.7	30.7	49.1

Table 2: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 200$.

Power correlation: $\sigma_{i,j} = 0.9^{\lvert i-j \rvert}$								
	1-Boot	3-Aug	3-gH	3-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.4	5.4	2.0	5.5	5.4	2.7	5.4	50.0
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_i = 0.25$ (equal-spaced)								
Control	5.3	3.6	1.8	5.0	5.1	2.6	5.2	49.0
Rejected	3.7	5.5	4.1	5.3	3.7	2.7	3.7	8.3
Ten $\theta_i = 0.25$ (in blocks)								
Control	5.3	3.6	1.4	4.8	5.0	2.2	5.1	48.3
Rejected	3.7	5.4	4.3	5.3	3.7	2.7	3.7	8.3
Twenty-five $\theta_i = 0.25$ (equal-spaced)								
Control	5.2	1.9	0.1	3.8	2.7	1.7	4.1	49.3
Rejected	9.5	11.4	9.7	13.6	10.1	7.8	10.8	22.6
Twenty-five $\theta_i = 0.25$ (in blocks)								
Control	4.8	2.2	0.1	4.1	3.4	1.4	4.4	47.3
Rejected	9.4	11.3	9.6	13.4	10.0	7.7	10.6	22.6
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	21.6	23.5	19.0	31.2	23.3	22.0	30.5	49.3

Table 3: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 200$.

Two-class structure: $\sigma_{i,j} = 0.5$ or -0.5								
	1-Boot	3-Aug	3-gH	3-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	4.9	4.9	1.0	4.8	4.9	3.0	4.9	50.8
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_i = 0.25$ (equal-spaced)								
Control	5.1	2.9	0.1	4.3	5.1	3.3	5.1	46.6
Rejected	3.1	4.9	3.9	5.5	3.1	2.6	3.1	7.4
Ten $\theta_i = 0.25$ (in blocks)								
Control	4.8	2.5	0.1	4.4	4.8	3.1	4.8	46.4
Rejected	3.1	4.9	3.9	5.5	3.1	2.6	3.1	7.5
Twenty-five $\theta_i = 0.25$ (equal-spaced)								
Control	3.5	1.3	0.3	2.5	1.8	1.1	2.8	44.6
Rejected	8.1	10.0	9.6	14.4	8.5	7.2	8.9	22.5
Twenty-five $\theta_i = 0.25$ (in blocks)								
Control	4.1	1.5	0.5	3.0	2.2	1.5	3.5	44.5
Rejected	8.1	10.0	9.6	14.4	8.5	7.1	8.9	22.5
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	17.4	19.4	19.1	35.0	19.0	20.1	36.1	49.7

Table 4: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 400$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 when all $\theta_i = 0$ and 2,000 for all other scenarios; and the number of bootstrap resamples is $B = 200$.

Common correlation: $\sigma_{i,j} = 0$								
	1-Boot	10-Aug	10-gH	10-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.3	5.3	0.0	1.7	5.3	5.1	5.3	55.1
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$								
Control	4.3	0.0	0.0	0.1	0.1	2.1	2.2	41.9
Rejected	11.0	19.9	28.0	59.5	11.8	14.0	29.5	68.7
Two hundred $\theta_i = 0.25$								
Control	2.7	0.0	0.0	0.4	0.0	0.0	0.4	29.9
Rejected	22.4	31.4	56.1	126.2	24.7	43.6	146.3	173.1
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	46.0	55.0	112.2	341.4	51.2	153.6	400.0	400.0
Common correlation: $\sigma_{i,j} = 0.5$								
	1-Boot	10-Aug	10-gH	10-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.5	5.5	0.1	5.5	5.5	2.1	5.5	51.9
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$								
Control	5.2	0.5	0.5	4.7	0.6	0.7	4.7	49.5
Rejected	18.2	27.0	29.1	47.2	19.9	17.4	33.4	84.5
Two hundred $\theta_i = 0.25$								
Control	3.6	0.7	0.6	4.7	0.5	1.2	4.7	51.5
Rejected	38.2	47.1	57.4	100.6	42.3	49.8	94.5	184.5
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	85.5	94.4	113.1	236.9	93.5	167.3	278.9	393.2

Table 5: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 400$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 when all $\theta_i = 0$ and 2,000 for all other scenarios; and the number of bootstrap resamples is $B = 200$.

Power correlation: $\sigma_{i,j} = 0.9^{\lvert i-j \rvert}$								
	1-Boot	10-Aug	10-gH	10-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.6	5.6	0.4	4.9	5.6	2.9	5.6	52.5
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$ (equal-spaced)								
Control	4.5	0.0	0.2	3.6	1.2	1.3	3.9	45.9
Rejected	14.6	23.4	27.7	49.1	15.9	14.3	24.9	71.8
One hundred $\theta_i = 0.25$ (in blocks)								
Control	4.6	0.0	0.2	3.8	1.7	1.5	4.2	45.3
Rejected	15.0	23.9	28.4	49.5	16.4	15.6	26.9	71.9
Two hundred $\theta_i = 0.25$ (equal-spaced)								
Control	3.9	0.0	0.0	1.6	0.0	0.0	1.8	42.6
Rejected	29.6	38.6	55.4	104.4	32.7	43.2	104.0	173.8
Two hundred $\theta_i = 0.25$ (in blocks)								
Control	3.4	0.0	0.1	2.6	0.3	0.4	2.8	43.5
Rejected	30.1	39.0	55.9	103.6	33.3	43.9	100.7	174.3
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	62.6	71.6	112.0	258.2	69.8	154.0	392.6	400.0

Table 6: Empirical FWEs and FDPs (in the rows 'Control') and average number of false hypotheses rejected (in the rows 'Rejected') for various methods, with $n = 100$ and $s = 400$. The nominal level is $\alpha = 5\%$, apart from the last column where it is $\alpha = 50\%$. The number of repetitions is 5,000 when all $\theta_i = 0$ and 2,000 for all other scenarios; and the number of bootstrap resamples is $B = 200$.

Two-class structure: $\sigma_{i,j} = 0.5$ or -0.5								
	1-Boot	10-Aug	10-gH	10-Boot	$\text{Aug}_{0.1}$	$\text{LR}_{0.1}$	$\text{Boot}_{0.1}$	$\text{Boot}_{0.1}^{Med}$
All $\theta_i = 0$								
Control	5.6	5.6	1.0	5.2	5.6	2.7	5.6	52.1
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
One hundred $\theta_i = 0.25$ (equal-spaced)								
Control	4.3	0.4	0.7	4.1	0.8	0.8	4.1	47.9
Rejected	15.8	24.6	28.0	47.1	17.1	14.7	27.3	81.4
One hundred $\theta_i = 0.25$ (in blocks)								
Control	5.3	0.7	1.1	5.4	1.1	1.3	5.4	47.6
Rejected	15.9	24.7	28.2	47.3	17.3	14.3	27.0	81.4
Two hundred $\theta_i = 0.25$ (equal-spaced)								
Control	4.3	0.2	0.3	3.2	0.2	0.3	3.2	44.3
Rejected	31.9	40.8	55.8	98.9	35.2	39.3	95.0	184.6
Two hundred $\theta_i = 0.25$ (in blocks)								
Control	3.9	0.2	0.2	3.4	0.3	0.4	3.3	44.7
Rejected	31.8	40.7	55.9	98.9	35.1	39.4	95.0	184.5
All $\theta_i = 0.25$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	67.1	76.1	112.2	237.0	74.9	142.0	349.9	398.9

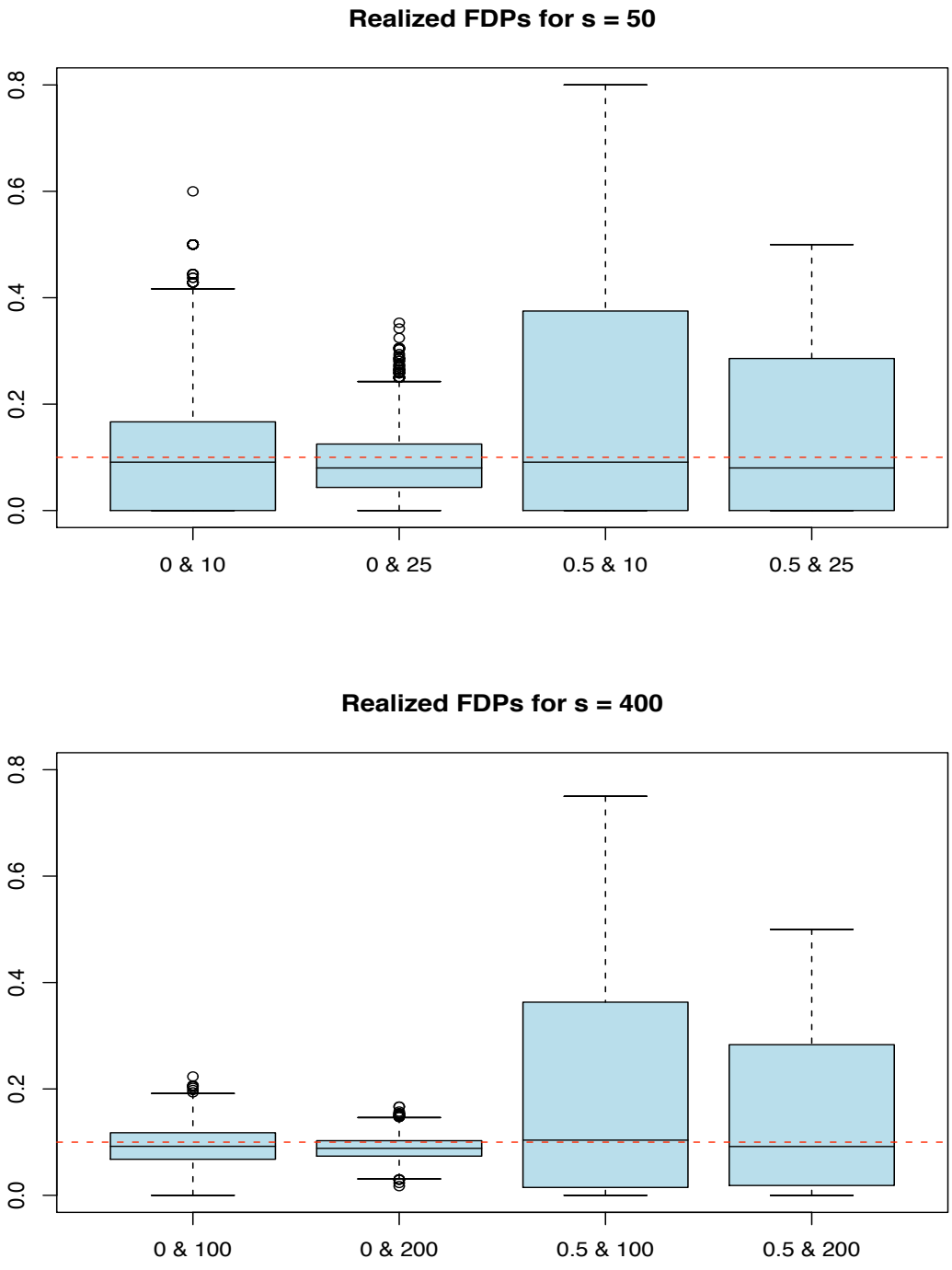

Figure 1: Boxplots of realized FDPs for various scenarios. The upper part shows scenarios 2, 3, 6, and 7 of Table 1; the lower part shows scenarios 2, 3, 6, and 7 of Table 4. For example, the name "0 & 10" in the upper part stands for "Common correlation = 0 & Ten $\theta_i = 0.25$". In both parts, the dashed line indicates $\gamma = 0.1$.